UNIVERZA V LJUBLJANI

FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Damir Možek

Analiza časovnih in prostorskih podatkov pri osebnih zavarovanjih

MAGISTRSKO DELO

Mentor: prof. dr. Igor Kononenko

Ljubljana, 2016

I Z J A V A O A V T O R S T V U

magistrskega dela

Spodaj podpisani/-a Damir Možek,

z vpisno številko 63960097,

sem avtor/-ica magistrskega dela z naslovom

Analiza časovnih in prostorskih podatkov pri osebnih zavarovanjih

S svojim podpisom zagotavljam, da:

- sem magistrsko delo izdelal/-a samostojno pod vodstvom mentorja (naziv, ime in priimek)

  prof. dr. Igor Kononenko

  in somentorstvom (naziv, ime in priimek)

- so elektronska oblika magistrskega dela, naslova (slov., angl.), povzetka (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko magistrskega dela

- in soglašam z javno objavo elektronske oblike magistrskega dela v zbirki »Dela FRI«.

V Ljubljani, dne 11.8.2016 Podpis avtorja/-ice: _____

ZAHVALA

Za usmerjanje in vzpodbujanje pri izdelavi magistrskega dela se iskreno zahvaljujem svojemu mentorju, prof. dr. Igorju Kononenku.
Zahvala gre tudi mojim bližnjim za podporo skozi leta študija.

Hvala vsem.

Damir

Kazalo vsebine

Kazalo slik

Kazalo tabel

Seznam uporabljenih kratic in simbolov

ARSO	Agencija republike Slovenije za okolje
DPZ	dodatno pokojninsko zavarovanje
GEOSS	Geometrično središče Slovenije
RMAE	relativna srednja absolutna napaka
RRMSE	koren relativne srednje kvadratne napake
SURS	Statistični urad republike Slovenije
SVM	metoda podpornih vektorjev
SZZ	Slovensko zavarovalno združenje
ZZZS	Zavod za zdravstveno zavarovanje Slovenije

Univerza v Ljubljani

Fakulteta za računalništvo in informatiko

Damir Možek

Analiza časovnih in prostorskih podatkov pri osebnih zavarovanjih

Povzetek

V tem magistrskem delu predstavljamo razvoj metodologije za analizo zavarovalniških podatkov. Ker so zavarovalniški podatki predstavljeni časovno in prostorsko, je za razvoj metodologije potreben poseben pristop. V ta namen uporabimo prijeme prostorsko-časovnega podatkovnega rudarjenja, ki nam omogočajo ustrezno obravnavo časovnih in prostorskih atributov.

Pri analizi podatkov se omejimo na podatke osebnih zavarovanj. V obravnavo zajamemo podatke o prijavah škodnih dogodkov na področju nezgod in bolezni. Zavarovalniške podatke navežemo na podatke o vremenskih razmerah v dnevu prijave odškodninskega zahtevka. S to navezavo želimo izkoristiti tezo o vplivu vremena na pojav nezgod. Nadejamo se, da bomo na ta način lažje napovedovali število odškodninskih zahtevkov in povprečno višino izplačanega odškodninskega zahtevka.

Najprej se lotimo reševanja klasifikacijskega problema. Uporabimo nekaj osnovnih klasifikacijskih algoritmov, vendar se stopnja uspešnosti napovedovanja pri vseh algoritmih izkaže za izredno nizko. Ker so problemi po naravi regresijski, se preizkusimo še v reševanju regresijskega problema. Tudi regresijski algoritmi ne dajo dosti boljših rezultatov. Preverimo ustreznost časovnega okna učne množice. Dobimo potrditev, da je časovno okno glede na podatke izbrano ustrezno. Nadalje preverimo, če instance obravnavajo ustrezno časovno skalo. Pridemo do sklepa, da je časovna skala izbrana ustrezno. Poskusimo lokalizirati problem tako, da podatke razdelimo. Pri različnih primerih razreza Slovenije pridemo do ugotovitve, da je ocena napovedi za vsako izmed lokalnih območij slabša od napovedi za celotno Slovenijo.

Postavljena metodologija se izkaže za delno uporabno. Uporabimo jo lahko za napovedovanje števila nezgod. S pomočjo diagnostike dobimo potrditev, da je za neuspeh kriva majhnost

množice obravnavanih dogodkov. Podamo predloge glede možnosti izboljšav. Nadejamo se, da se uporabnost postavljene metodologije pokaže v prihodnosti.

Za potrebe magistrskega dela pripravimo še zemljevide za spremljanje nezgod in bolezni v Sloveniji skozi čas. Na podlagi zemljevidov ugotavljamo trende gibanja za prihodnost. Identificiramo najpomembnejše dejavnike, odgovorne za nastanek nezgod. Na koncu naredimo še analizo vpliva nadmorske višine na pojav nezgod.

Ključne besede:

prostorsko-časovno podatkovno rudarjenje, strojno učenje, osebna zavarovanja, nezgode, vreme.

University of Ljubljana

Faculty of Computer and Information Science

Damir Možek

Analysing of temporal and spatial data in life insurance

Abstract

In this master's thesis we present the development of the methodology for the analysis of insurance data. Due to the fact that insurance data are presented temporally and spatially, a special approach is necessary for the development of the methodology. For this purpose we use approaches of spatio-temporal data mining, which enable us the appropriate treatment of temporal and spatial attributes.

When analysing data we limit ourselves on the data of personal insurance. Into the treatment we capture data on reports of the loss events in the field of accidents and diseases. Insurance data are linked to the data on weather conditions on the day of the report of the claim of compensation. By this linkage we wish to use the thesis on the influence of the weather on the accidents. We hope that in this way we shall predict the number of the claims of compensation and the average amount of the disbursed claims of compensation more easily.

Firstly, we deal with the solving of the classification problem. We use some of the basic classification algorithms, but the level of successfulness of predicting in case of all algorithms proves to be extremely low. Due to the fact that the problems are by nature regression, we try to solve the regression problem too. Even regression algorithms do not offer much better results. We check the adequacy of training set time window. We get the confirmation that the time windows is selected appropriately with respect to the data. Furthermore, we check if the instances deal with the appropriate time scale. We come to the conclusion that the time scale is selected appropriately. We try to localize the problem: we divide the data. In different cases of the cut of Slovenia we come to the ascertainment that the estimate of the prediction for each of the local areas is worse than for the entire Slovenia.

The set methodology proves to be partially useful. It can be used for predicting the number of accidents. By means of diagnostics we receive the confirmation that the failure is due to the smallness of the multitude of treated events. We give proposals regarding the possibility of improvements. We hope that the usefulness of the set methodology will become evident in the future.

For the needs of the master's thesis we also prepare the maps for following the accidents and diseases in Slovenia through time. On the basis of the maps we ascertain the trends for the future. We identify the most important factors, responsible for the emergence of the accidents. At the end we perform the analysis of the influence of the altitude for the emergence of the accidents.

Keywords:

spatio-temporal data mining, machine learning, personal insurance, accidents, weather.

1. Uvod

Kot predstavnice finančnih institucij so zavarovalnice dolžne skrbnega ravnanja ob zagotavljanju skladnosti poslovanja z zakonodajo. Poleg tega so primorane slediti trendom, če le želijo ohranjati svojo konkurenčnost na trgu. Želja po konkurenčnosti je vodilo, ki zavarovalnicam narekuje stalne izboljšave in prilagoditve procesov. Tu gre iskati vzrok, zakaj se v zavarovalnicah vedno kaj spreminja, se prilagaja in se išče nove, inovativne pristope.

1.1. Opredelitev problema

Na slovenskem so zavarovalnice prisotne že dalj časa in se uspešno spopadajo z izzivi informacijske družbe. V vseh teh letih so nakopičile že zajetne količine podatkov, ki lahko pridejo še zelo prav za izboljšanje samega poslovanja. Zagotovo je smiselno, da zavarovalnice te pridobljene podatke skušajo nekako uporabiti za povečanje lastne konkurenčnosti, vendar je za to potrebno iz podatkov znati izluščiti uporabne informacije. Takemu sistematičnemu iskanju informacij iz podatkov pravimo podatkovno rudarjenje.

Iz podatkov je mogoče dobiti celo vrsto vzorcev in pravil, ki nam nadalje služijo za optimizacijo in izboljševanje obstoječih procesov. Pri podatkovnem rudarjenju gre običajno za pregledovanje ogromnih količin podatkov, kar bi brez ustreznih orodij bilo težko izvedljivo. Pri tem so nam v veliko pomoč orodja za strojno učenje. S pomočjo teh orodij preizkušamo različne tehnike podatkovnega rudarjenja in nad podatki zaganjamo vnaprej pripravljene algoritme. Med orodji za podatkovno rudarjenje omenimo zgolj dva izmed najbolj uporabljanih: Weka [22] in Orange [28].

V zadnjem času v zavarovalništvu tehnike podatkovnega rudarjenja uporabljajo predvsem pri zaznavanju potencialnih prevar. Zavarovalnice so kot predstavnice finančnih ustanov vsekakor zaželena tarča goljufov. Podrobnejši pregled finančnih prevar v zavarovalniškem sektorju s podanimi rešitvami sta povzela Lookman in Balasubramanian [10]. Za slovenski prostor se je tega pri nas lotila Valand [20]. Žal se podatkovno rudarjenje v zavarovalništvu dogaja predvsem na področju neživljenjskih oziroma premoženjskih zavarovanj. Življenjska zavarovanja so tu nekoliko zapostavljena, oziroma je tam podatkovno rudarjenje trenutno prisotno v bistveno manjši meri. Po svoje je to sicer razumljivo, saj je trg življenjskih zavarovanj pri nas še vedno precej manjši od trga premoženjskih zavarovanj. Trend se je sicer tudi pri nas že obračal, a je

zaradi globalne gospodarske krize prodaja osebnih zavarovanj v zadnjem času spet nekoliko zamrla.

Kako uporabiti pridobljene podatke za izboljšanje procesov življenjskih zavarovanj? Na uspešnost poslovanja zavarovalnice imajo pomemben vpliv kazalniki o višini izplačanih škodnih zahtevkov. Poleg tega nam veliko pove tudi število odškodninskih zahtevkov. S pomočjo tehnik podatkovnega rudarjenja skušamo napovedati tudi takšne podatke. Bolj kot so napovedi točne, bolje se zavarovalnica odzove na razmere na trgu. Za napoved je potrebno spremljati odškodninske zahtevke skozi različna obdobja, glede na različne lokacije. Najbolje se tega lotevamo z metodami tako imenovanega časovno prostorskega podatkovnega rudarjenja [21].

Podatki, ki jih zajemajo na zavarovalnicah, imajo časovne in prostorske porazdelitve. V magistrskem delu analiziramo pristope podatkovnega rudarjenja z analizo prostorskih in časovnih podatkov. Podatke preuredimo v primerno obliko za napovedovanje odškodninskih zahtevkov. Napovedujemo s pomočjo klasifikacije kakor tudi regresije.

Za razumevanje spremljanja odškodninskih zahtevkov je pomembno razumevanje dejavnikov, ki vplivajo na njihov nastanek. Vpliv vremena na poškodbe je že dolgo znan in večkrat analiziran. Poškodbe kolka zaradi vpliva vremena so že v 1998 raziskovali Levy, Bensimon, Mayo in Leighton [9]. Poškodbe pri igranju ameriškega nogometa zaradi vplivov vremena sta analizirala Orchard in Powell [11]. Potrditev o vplivu vremena na poškodbe pa najdemo tudi v napotkih za varstvo in zdravje londonskega inštituta IET [25]. Teza o vplivu vremena na odškodninske zahtevke deluje obetavno. Zajeli smo podatke o vremenu na dan poškodbe in preizkusili njihovo uporabnost za napovedi.

1.2. Namen in cilji raziskave

Namen magistrske naloge je na realnih podatkih analizirati vpliv vremena na odškodninske zahtevke, prijavljene v zavarovalnici. Pridobljene informacije lahko zavarovalnica izkoristi za povečanje konkurenčnosti, obenem pa zniža stopnjo tveganja neuspeha.

Cilj magistrske naloge je na podlagi tehnik podatkovnega rudarjenja razviti metodologijo, s katero bo možno odškodninske zahtevke modelirati in napovedovati tako številčno kot tudi v smislu zneska. Poskušali smo poiskati povezave med odškodninskimi zahtevki iz preteklih obdobij in vremenskimi razmerami v času dogodkov. Na podlagi tega smo razvili metodologijo za napovedovanje odškodninskih zahtevkov.

Glavni prispevek magistrske naloge je razvita metodologija za zmožnost lokalnega modeliranja zavarovalniških podatkov in napovedovanja odškodninskih zahtevkov, ki je verificirana na realnih podatkih. Prostorska lokalnost napovedi je tu ključna predvsem zaradi raznolikosti populacij, zastopanih na posamezni lokaciji. Prav tako je pomembna časovna os, saj se vremenski vpliv spreminja skladno z letnimi časi.

V sklopu magistrskega dela smo za večletno časovno obdobje izdelali zemljevide pojavljanja nezgod v Sloveniji. Spremljali smo pojav nezgod in ugotavljali trende.

Analizirali smo vpliv posameznih dejavnikov na napovedovanje nezgod. Pri tem smo predstavili najvplivnejše dejavnike in njihov vpliv utemeljili z razlago.

Kot zanimivost smo preverili tudi vpliv nadmorske višine na odškodninske zahtevke. Poskusili smo poiskati vzorce in pravila v podatkih.

1.3. Metodološki pristop

Magistrska naloga obsega postavitev metodologije za analizo odškodninskih zahtevkov in je sestavljena iz teoretičnega in empiričnega dela.

V teoretičnem delu magistrske naloge smo uporabili metodo teoretičnega raziskovanja oziroma znanstvene deskripcije, ki obsega zbiranje in urejanje obstoječih dognanj, primerjavo ter interpretacijo le-teh.

V empirični delu magistrske naloge, ki obsega postavitev metodologije, smo uporabili splošno raziskovalno metodo spoznavnega procesa, ki obsega analizo, sintezo spoznanj ter zaključno sintezo novih spoznanj.
Podatke za verificiranje razvite metodologije smo pridobili od ene izmed večjih slovenskih zavarovalnic. Za potrebe analize smo nato zavarovalniške podatke povezali s podatki o vremenu. Podatki o vremenu v Sloveniji za pretekla leta so javno dostopni in smo jih pridobili na vremenskem portalu [26].

1.4. Omejitve in predpostavke pri raziskavi

Podatki, uporabljeni za verifikacijo metodologije, pripadajo eni od večjih slovenskih zavarovalnic. Zavoljo varovanja poslovnih skrivnosti so v magistrskem delu občutljivi podatki zakriti ali ustrezno zakodirani.

Izmenjava informacij med zavarovalnicami je na tem področju skromna. Tu gre za občutljive podatke, ki jih finančne institucije med seboj nerade delijo, tako da se raziskave na tem področju tipično ne objavljajo. Če že zasledimo kakšno publikacijo, so podatki običajno zakodirani in vprašanje je, koliko so kot takšni sploh uporabni. S tem izzivom se mora trenutno vsaka zavarovalnica spopadati bolj ali manj sama.

Pri analizi smo se omejili na zavarovalniške podatke osebnih zavarovanj. Poleg tega smo obravnavali le dva vzroka prijav iz osebnih zavarovanj. Zajeli smo le nezgode in bolezni; vsi ostali vzroki so v tej raziskavi izpuščeni. Zaradi obsega magistrske naloge in specifičnosti obravnave posameznih vzrokov je omejitev na manjše število vzrokov žal nujno potrebna.

Zaradi raznolikosti med prijavami skozi leta smo pri napovedih kot okno za pregled preteklih dogodkov zajeli le dogodke z omejenim časovnim pogledom nazaj. Skušali smo uporabiti podatke od vključno leta 2000 naprej. V primeru bistvenega odstopanja podatkov od povprečnega stanja zadnjih nekaj let smo uporabili krajši časovni interval. Podali smo primerjavo uspešnosti napovedi glede na različne časovne skale. Pri tem smo uporabili dnevno, tedensko in mesečno časovno skalo.

Pri spremljanju vremenskih podatkov smo se omejili na vremenske postaje, na katerih potekajo meritve vseh spremenljivk, o katerih poročajo na vremenskem portalu [26]. Zaradi te omejitve padavinske vremenske postaje v naše magistrsko delo niso zajete.

1.5. Struktura magistrskega dela

Magistrsko delo je razdeljeno v sedem poglavij. Uvodnemu poglavju sledi poglavje, kjer na kratko umestimo in predstavimo osebna zavarovanja. Posebej izpostavljene so vrste zavarovanj, zaobsežene v tem magistrskem delu. V tretjem poglavju zasledimo najprej nekaj na splošno o vremenu, nato so podrobneje predstavljene vremenske spremenljivke, uporabljene pri meritvah. Četrto poglavje govori najprej na splošno o podatkovnem rudarjenju, nato sledi predstavitev prostorsko-časovnega podatkovnega rudarjenja. Peto poglavje predstavlja osrednji del, v katerem je predstavljen razvoj metodologije. V šestem poglavju so podane sklepne ugotovitve, v katerih pregledamo izpolnjenost zastavljenih ciljev. Sedmo poglavje predstavlja zaključno poglavje, v katerem so podani zaključki in smernice nadaljnjega dela.

2. Osebna zavarovanja

Glede na podatke slovenskega zavarovalnega združenja (v nadaljevanju SZZ) poznamo dve večji skupini zavarovanj: osebna zavarovanja in premoženjska zavarovanja (opredeljeno v [30]). Osebnim zavarovanjem pravimo tudi življenjska zavarovanja. Nasprotno temu premoženjska zavarovanja sicer pojmujemo tudi kot neživljenjska zavarovanja. V tem magistrskem delu se osredotočamo na analizo nekaj vrst zavarovanj iz skupine osebnih zavarovanj.

SZZ [30] opredeljuje osebna zavarovanja kot zavarovanja, pri katerih so predmet zavarovanja osebe oziroma njihove osebne dobrine. Pod tem pojmujemo življenje, zdravje, delovno sposobnost in podobno. So zavarovanja, s katerimi si zagotovimo največjo mero varnosti za nepredvidene življenjske dogodke.

2.1. Vrste osebnih zavarovanj

SZZ nadalje osebna zavarovanja deli na več vrst. Vrste osebnih zavarovanj predstavljamo v nadaljevanju.

Zavarovanja pri nas razmestimo na več nivojih. Podrobnejšo razmestitev je v svojem delu pripravila Kastelic [6]. Mi za svoje potrebe razmestitev povzemamo po SZZ [30].

Vrste osebnih zavarovanj po SZZ:

1. življenjska zavarovanja,
2. nezgodna zavarovanja,
3. zdravstvena zavarovanja,
4. pokojninska zavarovanja.

2.1.1. Življenjsko zavarovanje

Življenjsko zavarovanje je zavarovanje, pri katerem želimo v primeru svoje smrti nekomu zagotoviti finančno varnost, hkrati pa želimo tudi sami varčevati za prihodnost. Poznamo različne oblike življenjskih zavarovanj, ki jih je možno nadgrajevati s priključevanjem dodatnih zavarovanj. Tudi življenjska zavarovanja razdelimo na več vrst. Povzeto po SZZ [30] govorimo o naslednjih vrstah življenjskih zavarovanj:

1. življenjsko zavarovanje za primer smrti.

   Možnih je več oblik tega zavarovanja:

   - zavarovanje za primer smrti za vse življenje, pogosto imenovano kot vseživljenjsko zavarovanje. Pri tem zavarovanju upravičenec dobi denar v primeru smrti zavarovane osebe.
   - Časovno omejeno zavarovanje za primer smrti, pogosto imenovano kot rizično zavarovanje. V tem primeru se denar izplača le, če zavarovana oseba umre v vnaprej dogovorjenem času trajanja zavarovanja.
   - Zavarovanje za primer smrti s padajočo zavarovalno vsoto, pogosto imenovano kot zavarovanje kreditojemalca. Posebnost tega zavarovanja je, da se zavarovalna vsota znižuje s preostalo zavarovalno dobo.

2. Življenjsko zavarovanje za primer smrti in doživetja.

   Ta vrsta zavarovanja je pogosto imenovana mešano življenjsko zavarovanje. Združuje zavarovanje za primer smrti in je hkrati tudi varčevanje. Ob poteku zavarovalne dobe se izplača dogovorjena zavarovalna vsota s pripisom dobička. V primeru smrti zavarovane osebe med trajanjem zavarovanja se upravičencu izplača zavarovalna vsota za primer smrti.

3. Življenjsko zavarovanje za primer doživetja.

   Pri tej vrsti zavarovanj se zavarovalna vsota in pripisani dobiček izplača le v primeru, če zavarovana oseba doživi dogovorjeno zavarovalno dobo.

4. Naložbeno življenjsko zavarovanje.

   Vrsta zavarovanja, pri kateri gre za zavarovanje v primeru smrti in hkrati varčevanje v investicijskih skladih ali drugih oblikah naložb.

K osnovnemu zavarovanju je mogoče priključiti še razna dodatna zavarovanja, s katerimi nadgrajujemo svojo varnost glede na potrebe. Možno je recimo skleniti dodatno nezgodno zavarovanje za primer smrti in trajne invalidnosti, zavarovanje za primer obolelosti za kritično boleznijo, dodatno zavarovanje za primer brezposelnosti, dodatno nezgodno zavarovanje otroka ter še celo vrsto drugih dodatnih zavarovanj. Odvisno je od zavarovalnice, kaj nudi iz svojega repertoarja.

Odškodninski zahtevki življenjskih zavarovanj obsegajo tudi vzroke prijav, ki jih obravnavamo v tem magistrskem delu. Pokrivajo tako nezgode kot tudi bolezni.

2.1.2. Nezgodna zavarovanja

Nezgodna zavarovanja zagotavljajo socialno varnost posamezniku in njegovi družini. Pokrivajo zavarovanje za primer smrti in invalidnosti; lahko imajo tudi dodatna kritja.

Tudi pri nezgodnih zavarovanjih imamo odškodninske zahtevke z nam ustreznimi vzroki prijav. Že samo ime nam pove, da ta vrsta zavarovanj pokriva prijavo nezgode. Zato so tudi zavarovanja te vrste obravnavana v našem magistrskem delu.

2.1.3. Zdravstvena zavarovanja

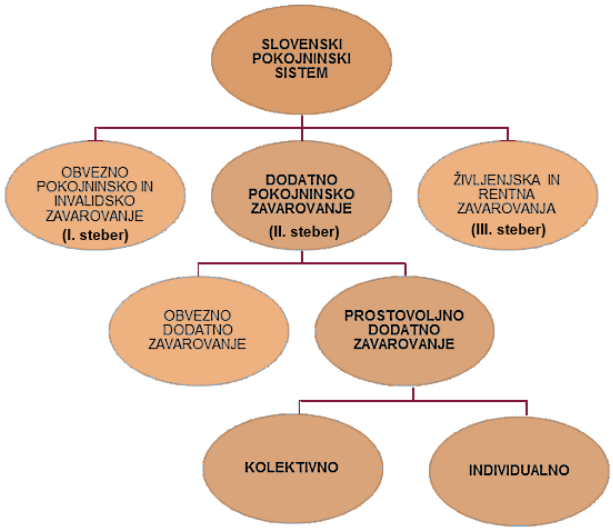
Obvezno zdravstveno zavarovanje imamo običajno sklenjeno pri zavodu za zdravstveno zavarovanje Slovenije (ZZZS). Ker pa to ne krije plačila zdravstvenih storitev v celoti, večina ljudi sklene tudi tako imenovano dopolnilno zdravstveno zavarovanje.

Dopolnilno zdravstveno zavarovanje je prostovoljno zavarovanje. Zavarovanje krije razliko med ceno celotne zdravstvene storitve in tistim delom, ki ga krije obvezno zdravstveno zavarovanje. Vrednost, do katere je ta storitev krita, je določena z višino najvišje zavarovalne vsote.

Zavarovalnice, ki v Sloveniji ponujajo dodatna zdravstvena zavarovanja, poslujejo kot samostojne družbe, ki ne ponujajo drugih vrst zavarovanja. Zavarovanja te vrste v našem magistrskem delu zato niso bila zaobsežena.

2.1.4. Pokojninska zavarovanja

V Sloveniji imamo za pokojninska zavarovanja tako imenovani sistem treh stebrov, ki ga predstavljamo na sliki1.



Slika 1: Sistem pokojninskega zavarovanja v Sloveniji (Vir: SZZ [30]).

Prvi steber deluje po načelu vzajemnosti in medgeneracijske solidarnosti. Določa pravico do pokojnine na podlagi plačanih prispevkov.

Drugi steber obsega obvezno dodatno pokojninsko zavarovanje, ki je nadomestilo za beneficirano delovno dobo in ga plačuje delodajalec. Poleg tega drugi steber obsega tudi dodatno pokojninsko zavarovanje (v nadaljevanju DPZ). Namenjen je tistim, ki že imajo sklenjen prvi steber, zbrana sredstva pa zavarovanci ob upokojitvi dobivajo v obliki dodatne pokojnine. Poznamo dve vrsti DPZ: individualno in kolektivno. Razlika med njimi je v tem, kdo vplačuje premijo. Pri individualnih premijo v celoti vplačuje posameznik, pri kolektivih premijo vsaj deloma vplačuje delodajalec.

Tretji steber predstavljajo razne druge oblike zavarovanj oziroma rentnih varčevanj. Bistvo je, da povečujejo socialno varnost.

Za zavarovalnice sta aktualna samo drugi in tretji steber. V primeru nezgodne smrti ali bolezni tudi pri teh zavarovanjih pride do izplačila odškodnine, tako da so tudi te vrste zavarovanj zaobsežene v našem magistrskem delu.

2.2. Obravnavani vzroki prijav

Kot smo napisali že v uvodu, smo se v magistrskem delu omejili zgolj na dva vzroka prijav. Obravnavali smo nezgode in bolezni. Poleg teh dveh vzrokov prijav poznamo v zavarovalnicah še vrsto drugih vzrokov prijav. Če naštejemo nekaj najpogostejših, poznamo tu še odkup, predujem, upokojitev in prenos sredstev. Večina izmed teh vrst prijav je tako specifična, da bi za njih bila potrebna posebna obravnava in zato niso zaobsežene v tem magistrskem delu.

Razdelajmo podrobneje, kaj vse obsega vzrok prijave tipa nezgoda. V skupino nezgod združujemo zdrse, padce, zlome, izpahe, nezgodne smrti, udarce, ugrize, ureznine, vbode, zastrupitve, opekline, odrgnine, natege in še vrsto drugih poškodb. Načeloma je vsem tem dogodkom skupno to, da so posledica nezgode, ki se je pripetila.

Drugi obravnavani tip vzroka prijav v magistrskem delu so bolezni. V skupino bolezni uvrščamo infarkte, kapi, razne infekcije, razne vrste raka in vrsto drugih bolezni, ki imajo običajno za posledico smrt zavarovanca.

Pri primerjavi obeh tukaj obravnavanih vzrokov lahko že takoj opozorimo na dejstvo, da je datum nastanka pri prijavi bolezni včasih nemogoče določiti. Ne poznamo vedno natančnega

dneva, kdaj smo zboleli za neko boleznijo. V takšnih primerih se poda zgolj približna ocena tega datuma. Pri boleznih je v dosti primerih dan nastanka opredeljen precej nenatančno. Že takoj na tem mestu povemo, da je to vzrok za bistveno manjši vpliv vremena pri vzroku bolezni.

2.3. Zajem podatkov škodnega dogodka

V primeru odškodninskega zahtevka je najprej potrebno urediti prijavo škodnega dogodka pri zavarovalnici. Škodne dogodke je sicer možno prijavljati tudi za nazaj, vendar je v interesu zavarovanca in tudi zavarovalnice, da se to opravi čimprej. Na ta način zavarovanec prej pride do odškodnine, zavarovalnica pa lažje spremlja poslovanje. Morebitna prijava škodnega dogodka za nazaj nam pri postavljanju metodologije predstavlja oviro. To pomeni, da bo ob prijavah za nazaj ob vnovičnem zajemu podatkov število pridobljenih škodnih dogodkov nekoliko večje.

Ob prijavi se v sistem zavarovalnice vnese osnovne podatke o dogodku. V določenih primerih se zahteva morebitna dodatna dokumentacija, lahko pa se zavarovanca napoti na dodatni pregled pri zdravniku. V primeru obravnavane zavarovalnice se je izkazalo, da zajem podatkov o kraju dogodka ni najbolje zastavljen. Namesto da bi se kraj dogodka izbiral iz spustnega menija, se ime kraja vnaša opisno. Zaradi tega pride dostikrat do vnosa nepravilnega imena kraja. Morda se v imenu izpusti kakšna črka, se ime napačno okrajša, ali pa se ime zasiči z dodatnim opisom. To precej oteži nadaljnjo obdelavo zajetih podatkov.

Najprej je bilo potrebno razviti algoritem, ki iz opisnega polja za kraj dogodka izlušči dejansko ime kraja, kjer se je dogodek pripetil. Ker obravnavamo samo področje Slovenije, so bili izločeni kraji, ki so izven tega področja. V veliko pomoč pri razreševanju nejasnosti pri tem je bila poštna številka. Poštna številka je tudi eden od podatkov, ki se zajemajo. Na podlagi poštne številke je bilo med drugim možno razlikovanje med kraji z identičnimi imeni. Razjasniti je bilo potrebno uporabo pogosto uporabljanih pojmov, večkrat se namesto domačega naslova vnese kar »doma« ali kaj podobnega. Svojevrstno težavo so predstavljale tudi tipkarske napake, ki jih je bilo potrebno vsaj nekako zaznati, če že ne odstraniti. Razviti algoritem za problem identifikacije kraja dogodka se je izkazal za učinkovitega in zato primernega za uporabo pri razvoju metodologije.

3. Vreme in spremenljivke

Vreme je meteorološko-klimatski izraz za stanje atmosfere, ki nastane pod vplivi vseh pomembnejših meteoroloških elementov in atmosferskih pojavov (temperatura, vlaga, zračni tlak, ...) [1]. Vreme lahko podrobneje opišemo z vremenskimi spremenljivkami, ki jih beležijo vremenske postaje.

3.1. Vremenske postaje

Povzeto po Seme [17]: glede na mobilnost vremenske postaje delimo v dve skupini. Prva skupina so stacionarne vremenske postaje, katerih lokacija se ne spreminja. V večini primerov gre tu za avtomatske vremenske postaje, ki izmerjene podatke samodejno pošiljajo v meteorološke centre. V drugo skupino vremenskih postaj spadajo postaje, ki se jim lokacija spreminja. Sem spadajo vremenske postaje, ki so pritrjene na balone in jih meteorologi dvakrat dnevno spuščajo v zrak.

V našem primeru smo se omejili na stacionarne vremenske postaje. Te nadalje delimo glede na namen in obseg programa dela. Poznamo:

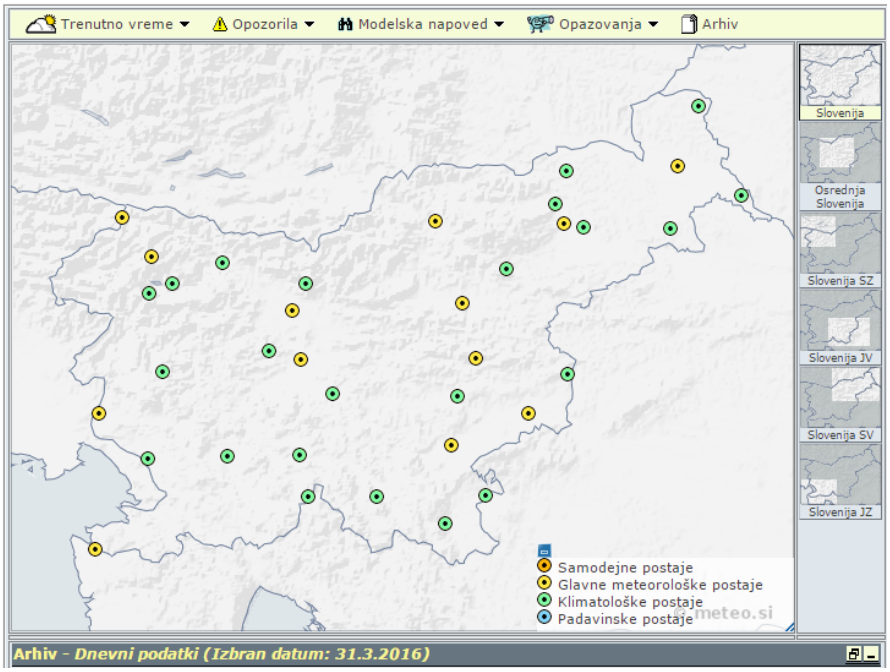
- glavne meteorološke ali sinoptične postaje,
- klimatološke ali podnebne postaje,
- padavinske postaje.

Glede na podatke državne meteorološke službe [26] je vrsta postaje odvisna od tega, katere meteorološke in biološke pojave in elemente tam opazujemo. Na sinoptičnih postajah potekajo meritve neprekinjeno in podatki se pošiljajo na vsake tri ure. Na teh postajah je največji nabor izvajanih meritev. Na podnebnih postajah se meritev opravlja trikrat dnevno. Merijo se enaki pojavi kot na sinoptičnih postajah. Merjenje na padavinski postaji se izvaja enkrat dnevno. Prav tako je na padavinskih postajah omejen nabor izvajanih meritev.

Za opazovanje glavnih meteoroloških spremenljivk smo v našem primeru upoštevali podatke sinoptičnih meteoroloških postaj in podnebnih postaj. Padavinske postaje smo zaradi omejenega nabora izvajanih meritev v našem primeru izpustili.

Trenutno število sinoptičnih in podnebnih postaj v letu 2016 v Sloveniji je 36 (podatek iz vremenskega portala [26]). Številka se skozi leta res bolj malo spreminja, se pa spremembe vseeno pojavljajo. V preteklosti so tako že vpeljevali nove postaje, ukinjali obstoječe ali pa

spremenili lokacijo obstoječe postaje. Na sliki 2 je predstavljena trenutna razporeditev meteoroloških postaj, ki jih obravnavamo v našem magistrskem delu.



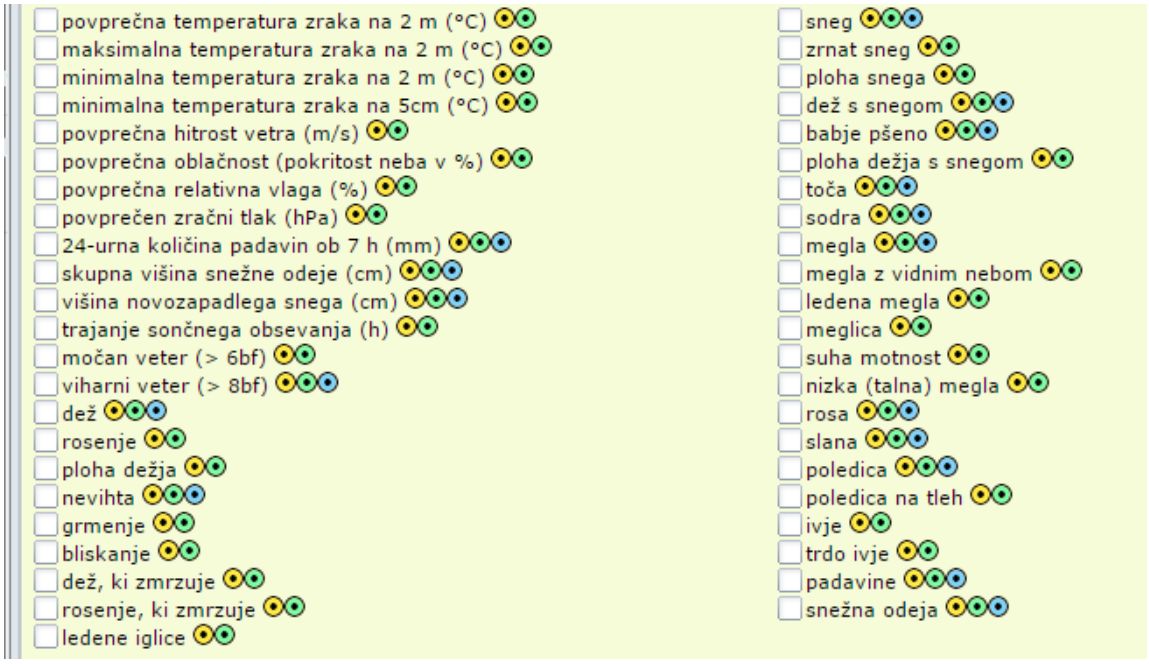
Slika 2: Sinoptične in podnebne postaje v Sloveniji leta 2016 (Vir: »Javne informacije Slovenije«, ARSO-met [26]).

3.2. Vremenske spremenljivke

Vreme natančneje opredeljujemo z vremenskimi spremenljivkami. Vremenske spremenljivke so pomembne za opis stanja ozračja in procesov, ki se odvijajo v njem. Pridobimo jih na vremenskih postajah z merjenjem meteoroloških in bioloških pojavov in elementov.

Meritve se izvajajo na meteoroloških postajah, od tam pa so podatki posredovani v meteorološke centre. Podatke o vremenskih spremenljivkah za pretekla obdobja je možno pridobiti na vremenskem portalu agencije republike Slovenije za okolje (v nadaljevanju ARSO). Vsi podatki na portalu ARSO so javno dostopni. Za potrebe magistrskega dela smo uporabili vse vremenske spremenljivke, ki so na portalu na voljo.

Na sliki 3 so predstavljene vremenske spremenljivke, ki jih je na portalu ARSO mogoče spremljati. Iz slike je razvidna tudi razlika med meritvami na postajah. Opaziti je, da na padavinskih postajah (modra barva) ne izvajajo vseh predstavljenih meritev. Padavinske postaje so prav zaradi tega razloga izpuščene iz nabora vremenskih postaj, upoštevanih v tem magistrskem delu.



Slika 3: Dnevne meteorološke spremenljivke (Vir: »Javne informacije Slovenije«, ARSO-met [26]).

3.3. Podatki o vremenu

Za analizo vpliva vremena je bilo v magistrskem delu potrebno pridobiti podatke o vremenu za različna opazovana področja. Podatke za iskano področje dobimo tako, da za vse kraje na opazovanem področju poiščemo najustreznejšo opazovalno postajo na iskani dan. Ta opazovalna postaja ni nujno tudi najbližja postaja, ampak poiščemo kraju bližnjo opazovalno postajo s podobno nadmorsko višino. Na ta način najdena opazovalna postaja naj bi precej dobro odražala vremenske parametre iskanega kraja. Za iskani kraj se privzame, da je na iskani dan vreme tam bilo enako vremenu, izmerjenemu v najdeni opazovalni postaji. Enak postopek uberemo za vse kraje, ki ležijo znotraj opazovanega področja. Vreme za celotno opazovano lokacijo potem določimo na osnovi povprečenja vremenskih spremenljivk vseh krajev opazovanega področja.

Na podlagi uporabljene časovne skale se v magistrski nalogi ustrezno povprečijo tudi podatki za izbrano obdobje. V primeru, ko je kot časovna skala izbran teden, se povprečijo podatki o vremenu za vse kraje za obdobje celotnega tedna, v primeru meseca pa za obdobje celotnega meseca. Uporaba časovnih skal je podrobneje razložena v nadaljevanju tega dela.

Precej vremenskih spremenljivk je podanih v logični obliki (true, false). Vrednosti teh spremenljivk so bile za potrebe magistrskega dela pretvorjene v numerično obliko. Vrednost je pretvorjena tako, da spremenljivka odraža delež vrednosti na opazovani lokaciji. V primeru, da

pojava na lokaciji ni, je to predstavljeno z deležem 0,0. Ko na pojav naletimo v vseh krajih, naraste ta delež na vrednost 1,0. Vrednosti, izražene z deleži, ohranijo več informacije in jim zato pripisujemo večji pomen.

4. Podatkovno rudarjenje

Podatkovno rudarjenje je širok pojem. Podajamo splošno definicijo podatkovnega rudarjenja, nato sledi navezava na prostorsko-časovno rudarjenje. Slednje je uporabljeno tudi v analizi tega magistrskega dela.

4.1. Definicija podatkovnega rudarjenja

Na uradni strani orodja za podatkovno rudarjenje Weka kot učno gradivo priporočajo knjigo »Data Mining: Practical Machine Learning Tools and Techniques« [22]. V tej knjigi zasledimo, da je podatkovno rudarjenje definirano kot proces odkrivanja vzorcev iz podatkov. Pri podatkovnem rudarjenju gre za ekstrakcijo implicitnih, prej neznanih in potencialno uporabnih informacij iz podatkov. Rudarjenje poteka avtomatsko ali vsaj polavtomatsko. Ideja je izgradnja računalniških programov, ki v podatkovnih bazah iščejo pravila in vzorce.

Vzemimo zdaj za primerjavo še opis iz knjige »Data Mining: Concepts and Techniques« [4]. Tu avtorji definirajo podatkovno rudarjenje kot ekstrakcijo oziroma rudarjenje znanja iz velikih količin podatkov. V nadaljevanju sicer tudi opozarjajo na neprimernost samega imena. Če izhajamo iz primerjave z rudarjenjem zlata, gre tam za rudarjenje zlata iz skal in peska, a vendar se rudarjenje zlata zaradi tega ne imenuje rudarjenje skal ali rudarjenje peska. Po tej logiki bi se podatkovno rudarjenje moralo imenovati rudarjenje znanja iz podatkov. Izraz je vsekakor predolg za uporabo. Krajša inačica – rudarjenje znanja, pa ne odraža rudarjenja iz ogromne količine podatkov. Zaradi tega se je uveljavil izraz podatkovno rudarjenje.

Po drugi strani avtorji v [4] navajajo tudi, da lahko na podatkovno rudarjenje gledamo kot rezultat naravne evolucije v informacijski tehnologiji. S pojavom zbiranja podatkov in mehanizmov za kreiranje podatkovnih baz so se pojavili predpogoji za razvoj učinkovitejših mehanizmov za shranjevanje in branje podatkov ter procesiranje transakcij. Podrobna analiza podatkov, kar podatkovno rudarjenje pooseblja, se je tako sama ponujala kot naslednji logični korak.

4.2. Prostorsko-časovno podatkovno rudarjenje

Z novimi tehnologijami se je v preteklem obdobju zajem časovnih in prostorskih podatkov občutno povečal. Prostorsko-časovno podatkovno rudarjenje se čedalje bolj uveljavlja kot odgovor na to povečanje.

Avtorji v [18] navajajo, da prostorsko-časovno rudarjenje preučuje proces odkrivanja zanimivih, prej nepoznanih, vendar potencialno uporabnih vzorcev, iz ogromnih podatkovnih baz s prostorskimi in časovnimi podatki. Prednost časovno-prostorskega rudarjenja je po Petelinu [13] ustrezna obravnava prostorskih in časovnih atributov.

V nadaljevanju podajamo opis in probleme prostorsko-časovnega rudarjenja podatkov, povzeto po Venkateswaru in ostalih [21].

Časovno-prostorski objekt je definiran kot objekt, ki ima vsaj eno časovno in eno prostorsko lastnost. Prostorsko-časovne množice obsegajo spreminjanje prostorskih vrednosti skozi čas. Najdemo jih na različnih področjih:

- meteorologija: vremenski podatki, tornadi, suše…
- biologija: gibanje živali, izumrtje vrst…
- poljedelstvo: žetev…
- gozdarstvo: gozdna rast, gozdni požari, sekanje gozdov, pogozdovanje…
- medicina: napredovanje raka…
- geografija: zgodovina potresov, aktivnosti vulkanov…
- ekologija: spremljanje onesnaženj…
- transport: nadzor prometa, načrtovanje prometa…

Modeliranje prostorsko-časovnih podatkov je problematično iz dveh razlogov. Prvi razlog je nenehno spreminjanje prostorsko-časovnih objektov. Drugi razlog je vpliv sosednjih objektov drug na drugega.

Pri nas se je s temo prostorsko-časovnega rudarjenja podatkov spopadel Petelin. Metodologijo na osnovi prostorsko-časovnega rudarjenja je uporabil v svoji doktorski disertaciji [13].

4.2.1. Prostorsko-časovno rudarjenja podatkov na primeru

Prostorsko-časovno podatkovno rudarjenje se je kot uporabno izkazalo v primeru preučevanja podnebnih sprememb. V nadaljevanju je na primeru predstavljeno, kako so se Ganguly in sodelavci [3] spoprijeli s tem problemom. Pokazalo se je, da z relativno preprostimi pristopom

podatkovnega rudarjenja dobimo vpogled v povsem nova znanstvena spoznanja. Podajamo krajši vpogled v pristop k temu projektu.

Opazovanje podnebja danes poteka preko raznih senzorjev, kot so sateliti, vremenski radarji in vrsta drugih senzorjev. Hitrost kopičenja teh podatkov žal veliko presega zmožnost analize le-teh. Pri tem gre za zajem prostorskih in časovnih podatkov, tako da pri analizi uporabimo tehnike prostorsko-časovnega podatkovnega rudarjenja.

Na primeru se je pokazalo, da analiza podnebnih podatkov predstavlja poseben izziv. Metodologije prostorsko-časovnega rudarjenja so zahtevale prilagoditve in celo razvoj novih pristopov.

Izzivi prostorsko-časovnega podatkovnega rudarjenja, ki jih je bilo potrebno upoštevati pri analizi podnebnih podatkov:

- zavedanje osnovnega zakona geografije: »Vse je povezano z vsem, vendar bližnje stvari bolj kot oddaljene.«
- Podatke klasificiramo ali regresiramo. Statistične napovedi so uporabne pri stabilnem podnebju, vendar ne pri spremenljivem podnebju.
- Detekcija lokalnih nestabilnosti, ki odstopajo od relativnih podatkov svojih sosed. Razločevanje teh nestabilnosti od meritvenih napak.
- Kategorizacija podnebnih režimov in določitev podnebnih indicev.
- Negotovost metodologije podatkovnega rudarjenja in tveganje, ki ga to prinaša.

Eden izmed rezultatov projekta je bil pokazatelj nevarnosti »afrikanizacije« Španije. Pojem predstavlja nevarnost naraščanja temperatur v kombinaciji s hudim pomanjkanjem padavin.

5. Razvoj metodologije

V tem razdelku podrobneje predstavljamo potek razvoja metodologije za napovedovanje odškodninskih zahtevkov.

5.1. Priprava podatkov

Eden prvih korakov je pridobitev podatkov, v okviru katerih želimo podatkovno rudarjenje izvajati. V našem primeru smo podatke o odškodninskih zahtevkih in njihovih izplačilih pridobili od ene izmed večjih slovenskih zavarovalnic. Te podatke smo navezali na podatke o vremenu v času škodnega dogodka. Podatki o vremenu so javno dostopni in pridobljeni iz vremenskega portala ARSO.

5.1.1. Določitev atributov

Glede vremenskih spremenljivk, ki smo jih uporabili za atribute, je bilo veliko napisanega že v razdelku 3.2. Naj samo ponovimo, da so bile za atribute uporabljene vse vremenske spremenljivke, ki so nam na voljo.

Poleg vremenskih spremenljivk je bilo dodanih še nekaj splošnih spremenljivk, ki natančneje opisujejo lastnosti glede na izbrano časovno skalo. Sem uvrščamo atribute, kot so številka meseca v letu, številka tedna v letu, ime dneva v tednu, delovnik, letni čas in lunina mena. Izbrana časovna skala natančneje predpisuje, kateri izmed atributov se pri kakšni skali uporabijo. Nima smisla, da pri mesečni časovni skali govorimo o lunini meni. Postopek za izbiro najbolj ustrezne časovne skale je podan v nadaljevanju.

Atributi lahko zavzamejo diskretne ali zvezne vrednosti. Če algoritmi zahtevajo atribute določene vrste, jih v ta namen prilagodimo. Več o tem je prikazano na primeru kasneje. Na tem mestu zgolj opozorimo, da smo diskretne vrednosti vremenskih atributov pretvorili v zvezne.

V tabeli 1 podajamo seznam vseh atributov, uporabljenih pri postavljanju metodologije. Pomen atributa je mogoče razbrati iz njegovega imena. Poleg imen atributov so podane njihove zaloge vrednosti in enote. Zadnji atribut je ciljni atribut in je v našem primeru imel različne zaloge vrednosti. Več o izbiri ciljnega atributa je napisano v naslednjem razdelku.

Tabela 1: Seznam atributov.

Ime atributa	**zaloga vrednosti**	**enota**
Mesec	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	
Teden	1 .. 53	
DanVTednu	ponedeljek, torek, sreda, četrtek, petek, sobota, nedelja	
Delovnik	delovni, nedelovni	
LuninaMena	prazna luna, prvi krajec, polna luna, zadnji krajec	
LetniCas	pomlad, poletje, jesen, zima	
PovprecnaTemperaturaZrakaNa2m	-10,9 .. 27,6	ºC
MinimalnaTemperaturaZrakaNa2m	-16,1 .. 20,2	ºC
MaksimalnaTemperaturaZrakaNa2m	-8,4 .. 37,5	ºC
MinimalnaTemperaturaZrakaNa5cm	-19,6 .. 18,9	ºC
PovprecnaHitrostVetra	0,4 .. 4,2	m/s
PovprecnaOblacnost	0,3 .. 100	%
PovprecnaRelativnaVlaga	41 .. 97,9	%
PovprecenZracniTlak	941 .. 1002,7	hPa
DnevnaKolicinaPadavin	0 .. 83	mm
SkupnaVisinaSnezneOdeje	0 .. 320	cm
VisinaNovozapadlegaSnega	0 .. 50	cm
TrajanjeSoncnegaObsevanja	0 .. 14	h
MocanVeter(> 6bf)	0	%
ViharniVeter(> 8bf)	0 .. 54	%
Dez	0 .. 100	%
Rosenje	0 .. 61	%
PlohaDezja	0 .. 67	%
Nevihta	0 .. 87	%
Grmenje	0 .. 68	%
Bliskanje	0 .. 50	%
DezKiZmrzuje	0 .. 71	%
RosenjeKiZmrzuje	0 .. 24	%
LedeneIglice	0 .. 2	%
Sneg	0 .. 100	%
ZrnatSneg	0 .. 19	%
PlohaSnega	0 .. 36	%
DezSSnegom	0 .. 49	%
BabjePseno	0 .. 24	%
PlohaDezjaSSnegom	0 .. 11	%
Toca	0 .. 23	%
Sodra	0 .. 45	%
MeglaZVidnimNebom	0 .. 38	%
LedenaMegla	0 .. 2	%
Meglica	0 .. 54	%
SuhaMotnost	0 .. 12	%
TalnaMegla	0 .. 13	%
Rosa	0 .. 80	%
Slana	0 .. 88	%
Poledica	0 .. 21	%
Ivje	0 .. 54	%
TrdoIvje	0 .. 7	%
Padavine	0 .. 100	%
SneznaOdeja	0 .. 100	%
CILJNI ATRIBUT		

5.1.2. Določanje ciljnega atributa

Pri določanju ciljnega atributa se najprej vprašamo, kakšne so vrednosti naših ciljnih atributov: diskretne ali zvezne. V primeru diskretnih vrednosti gre za klasifikacijski problem; v primeru zveznih vrednosti govorimo o regresijskem problemu. Glede na vrsto problema je odvisno, katere metode se uporabijo za reševanje le-tega. Odločili smo se, da v našem primeru najprej poskusimo z reševanjem klasifikacijskega problema. To pomeni, da za ciljni atribut določimo diskretne vrednosti. Potem smo poskusili še z reševanjem regresijskega problema, kjer smo za ciljni atribut iskali zvezne vrednosti.

V našem primeru nas zanima številčno napovedovanje odškodninskih zahtevkov in tudi napovedovanje samega odškodninskega zneska. Pri tem gre za dva ločena problema, ki jih je zato ločeno potrebno tudi obravnavati. V prvem primeru nam ciljni atribut pove število prijavljenih odškodninskih zahtevkov, v drugem primeru povprečno izplačilo obravnavanih prijav odškodninskih zahtevkov. Vrednosti ciljnih atributov smo diskretizirali v pet enakomerno porazdeljenih skupin. O diskretizaciji bomo več povedali v nadaljevanju. V tabelah 2 in 3 podajamo zaloge vrednosti obeh primerov ciljnih atributov pri reševanju klasifikacijskega problema.

Tabela 2: Zaloge vrednosti ciljnega atributa število odškodninskih zahtevkov.

vrednost	NEZGODE (‰ zavarovanih oseb)	BOLEZNI (€/zavarovanec)
malo	0,000 .. 0,127	0,000 .. 0,014
nekaj	0,128 .. 0,151	0,015 .. 0,019
srednje	0,152 .. 0,175	0,020 .. 0,024
precej	0,176 .. 0,203	0,025 .. 0,030
veliko	>= 0,204	>= 0,031

Tabela 3: Zaloge vrednosti ciljnega atributa število povprečno izplačilo odškodnine.

vrednost	NEZGODE (‰ zavarovanih oseb)	BOLEZNI (€/zavarovanec)
mizerno	0,000 .. 0,052	0,000 .. 0,027
nizko	0,053 .. 0,074	0,028 .. 0,042
srednje	0,075 .. 0,096	0,043 .. 0,059
visoko	0,097 .. 0,125	0,060 .. 0,088
ekstremno	>= 0,126	>= 0,089

Normalizacija ciljnih atributov

Število zavarovancev, pri katerih je možno, da pride do prijave odškodninskega zahtevka, ni stalno in se v času spreminja. To pomeni, da je bilo potrebno podatke zaradi njihove primerljivosti normalizirati s številom zavarovancev v opazovanem trenutku. Normalizirano

število prijavljenih odškodninskih zahtevkov nam ponazarja **delež vseh prijavljenih zahtevkov**. Povprečna vrednost izplačila zahtevka pa po normalizaciji predstavlja **višino povprečnega izplačila na zavarovanca**.

5.1.3. Pregled in čiščenje podatkov

Po določitvi vseh atributov je potreben podrobnejši pregled podatkov. Pri pregledu skušamo odkriti potencialne težave in jih po potrebi odpraviti. Osredotočamo se na smiselnost vrednosti atributov, manjkajoče vrednosti in ostala morebitna odstopanja. Odkriti je potrebno morebitne šume v podatkih, ki bi negativno vplivali na rezultate. Manjkajoče vrednosti utegnejo predstavljati problem, vendar smo v našem primeru uporabili algoritme, ki znajo delati tudi z manjkajočimi vrednostmi.

V veliko pomoč pri pregledovanju so nam bila vizualizacijska orodja, ki jih ponuja orodje Weka [22]. S pomočjo pripomočkov smo zlahka odkrili odstopanja in jih raziskali. V veliko pomoč pri tem nam je bilo tudi orodje Excel, ki se je prav tako izkazalo za izredno priročno pri odkrivanju odstopanj.

Pri pregledu smo odkrili določena temperaturna odstopanja pri posameznih dneh. Pojavljale so se nerealne temperature zraka, ki so precej odstopale od preostalih temperatur. Pri dveh dnevih smo zasledili celo temperaturo do 152 ºC. Podatki so bili ustrezno prilagojeni in faza čiščenja je bila s tem pripeljana do konca.

5.2. Evalvacija atributov

Evalvacija ali ocena pomembnosti atributov sicer ni nujna, se pa izkaže za uporabno ob preveliki kompleksnosti modela. S pomočjo ocene atributov identificiramo nerelevantne atribute in jih odstranimo. S tem model poenostavimo in posledično pridobimo pri hitrosti vzorčenja.

Pri evalvaciji atributa gre za ocenitev vpliva tega atributa na ciljni atribut. Nižja, kot je pri tem ocena tega atributa, manj pomemben se izkaže atribut za določitev vrednosti ciljnega atributa. Atribute z zanemarljivim vplivom na določitev ciljnega atributa zavoljo poenostavljenja modela zavržemo. Odstranitev takšnih atributov na sam ciljni atribut ne bi smela imeti bistvenega vpliva. Ocena pomembnosti atributov nam omogoča dober vpogled v model.

Poznamo več evalvatorjev, s pomočjo katerih ocenimo relevantnost atributa. V našem primeru smo uporabili štiri izmed njih. Prva dva za oceno atributov klasifikacijskega problema in druga dva za oceno atributov regresijskega problema.

5.2.1. Razmerje informacijskega prispevka

Prvi izmed evalvatorjev, uporabljen pri reševanju klasifikacijskega problema, je bil razmerje informacijskega prispevka. Povzeto po [7] je problem pri informacijskem prispevku (angl. Information gain), da kvaliteta atributa s številom vrednosti atributa kvečjemu raste. Zato je bilo definirano razmerje informacijskega prispevka (angl. gain-ratio):

$$GainR(A) = \frac{Gain(A)}{H_A}$$

Omenjeni problem informacijskega prispevka je tu odpravljen z normalizacijo informacijskega prispevka z entropijo vrednosti atributa.

Vpeljava notacije:

n – število učnih primerov,

$n_{k.}$ – število učnih primerov iz razreda r_k,

$n_{.j}$ – število učnih primerov z j-to vrednostjo danega atributa A,

n_{kj} – število učnih primerov iz razreda r_k in z j-to vrednostjo danega atributa A.

Vpeljava aproksimacije verjetnosti iz učne množice primerov:

$p_{kj} = n_{kj}/n$,

$p_{k.} = n_{k.}/n$,

$p_{.j} = n_{.j}/n$,

$p_{k|j} = p_{kj}/p_{.j} = n_{kj}/n_{.j}$

Vpeljava entropije:

H_R - entropija razredov:

$$H_R = -\sum_k p_{k.} \log p_{k.}$$

H_A - entropija vrednosti danega atributa:

$$H_A = -\sum_j p_{.j} \log p_{.j}$$

H_{RA} - entropija produkta dogodkov razred-vrednost atributa:

$$H_{RA} = -\sum_k \sum_j p_{kj} \log p_{kj}$$

$H_{R|A}$ – pogojna entropija razreda pri dani vrednosti atributa:

$$H_{R|A} = H_{RA} - H_A$$

$Gain(A)$ – informacijski prispevek:

$$Gain(A) = H_R + H_A - H_{RA} = H_R - H_{R|A}$$

Rezultati ocene atributov za primer nezgod so podani v tabeli 4. Oceno atributov za primer bolezni najdemo v dodatku A.1. V oceno so zajeti podatki od leta 2010 do konca leta 2015. Atributi, ocenjeni kot nepomembni, so v tabeli izpuščeni; ostali atributi so predstavljeni v padajočem vrstnem redu glede na njihovo pomembnost.

Tabela 4: Ocena atributov z razmerjem informacijskega prispevka pri nezgodah.

Število nezgod		**Povprečno izplačilo nezgode**	
atribut	**ocena**	**atribut**	**ocena**
Rosa	0,06630	SkupnaVisinaSnezneOdeje	0,07833
TrajanjeSoncnegaObsevanja	0,06008	Dez	0,04396
PovprecnaOblacnost	0,05983	SneznaOdeja	0,03489
Mesec	0,05053	MaksimalnaTemperaturaZrakaNa2m	0,03341
Teden	0,04753	TrajanjeSoncnegaObsevanja	0,03311
PovprecnaRelativnaVlaga	0,04634	Rosa	0,03109
Dez	0,04598	PovprecnaTemperaturaZrakaNa2m	0,03022
MaksimalnaTemperaturaZrakaNa2m	0,04146	Mesec	0,02967
Rosenje	0,03924	Rosenje	0,02749
PovprecnaTemperaturaZrakaNa2m	0,03627	Teden	0,02723
PovprecnaHitrostVetra	0,03421	PovprecnaOblacnost	0,02639
SneznaOdeja	0,03222	PovprecnaRelativnaVlaga	0,02466
Meglica	0,03029	Padavine	0,02274
DezSSnegom	0,02966	Meglica	0,02115
SkupnaVisinaSnezneOdeje	0,02758	DnevnaKolicinaPadavin	0,02035
Padavine	0,02439	DanVTednu	0,01914
MinimalnaTemperaturaZrakaNa2m	0,02125	Delovnik	0,01086
DnevnaKolicinaPadavin	0,02114	LetniCas	0,00840
DanVTednu	0,02080	LuninaMena	0,00150
LetniCas	0,01181		
Delovnik	0,00847		
LuninaMena	0,00105		

5.2.2. ReliefF

Kot drugi evalvator pri reševanju klasifikacijskega problema smo uporabili algoritem ReliefF. Povzeto po [7] je algoritem ReliefF izboljšana inačica algoritma Relief, ki se uporablja za ocenjevanje atributov, močno odvisnih med seboj. Uporabili smo ga zato, ker se odvisnost atributov močno kaže tudi v našem primeru. Algoritem Relief za vsak učni primer poišče najbližji primer iz istega razreda in najbližji primer iz nasprotnega razreda ter na podlagi tega oceni kvaliteto atributa. Funkcijo algoritma Relief opišemo z:

$$Relief(A_i) = \frac{\sum_j p_{.j}^2 * Gini'(A_i)}{\sum_k p_{k.}^2 (1 - \sum_k p_k^2)} = konstanta * \sum_j p_{.j}^2 * Gini'(A_i)$$

pri čemer velja

$$Gini'(A) = \sum_j \left(\frac{p_{.j}^2}{\sum_j p_{.j}^2} * \sum_k p_{k|j}^2 \right) - \sum_k p_{k.}^2$$

Preostali simboli formul so pojasnjeni v razdelku 5.2.1.

Algoritem ReliefF glede na [7] vsebuje naslednje razširitve:

- uporaba nepopolnih podatkov,
- iskanje k najbližjih zadetkov/pogreškov,
- reševanje večrazrednih problemov.

Rezultati ocene atributov z algoritmom ReliefF za primer nezgod so podani v tabeli 5, rezultate ocene za primer bolezni pa najdemo v dodatku A.2. V tabelah so prikazani zgolj atributi, ki so bili ocenjeni kot pomembni za določitev ciljnega atributa. Algoritem smo poganjali toliko časa, dokler niso v naboru atributov ostali samo atributi, ki imajo vpliv na izbiro ciljnega atributa. V tabelah so atributi razporejeni v padajočem vrstnem redu glede na njihovo pomembnost.

Tabela 5: Ocena atributov z ReliefF pri nezgodah.

Število nezgod		**Povprečno izplačilo nezgode**	
atribut	**ocena**	**atribut**	**ocena**
DanVTednu	0,03764	PovprecnaOblacnost	0,00381
LetniCas	0,01745	TrajanjeSoncnegaObsevanja	0,00363
LuninaMena	0,01704	MinimalnaTemperaturaZrakaNa2m	0,00340
Delovnik	0,01368	PovprecenZracniTlak	0,00338
VisinaNovozapadlegaSnega	0,01075	MinimalnaTemperaturaZrakaNa5cm	0,00310
SuhaMotnost	0,01071	PovprecnaRelativnaVlaga	0,00291
SkupnaVisinaSnezneOdeje	0,01056	DezSSnegom	0,00290
Meglica	0,00997	PovprecnaTemperaturaZrakaNa2m	0,00281
Mesec	0,00929	Sneg	0,00239
PlohaDezjaSSnegom	0,00914	DnevnaKolicinaPadavin	0,00209
TrajanjeSoncnegaObsevanja	0,00873	MaksimalnaTemperaturaZrakaNa2m	0,00196
ViharniVeter	0,00809	LedeneIglice	0,00157
PovprecnaOblacnost	0,00780	Rosenje	0,00117
Teden	0,00643	Dez	0,00115
PovprecnaTemperaturaZrakaNa2m	0,00512	ZrnatSneg	0,00072
MaksimalnaTemperaturaZrakaNa2m	0,00501	DezKiZmrzuje	0,00065
MinimalnaTemperaturaZrakaNa2m	0,00444	PlohaDezja	0,00063
MinimalnaTemperaturaZrakaNa5cm	0,00426	Nevihta	0,00052
PovprecnaRelativnaVlaga	0,00366	PlohaDezjaSSnegom	0,00040
PlohaSnega	0,00352	BabjePseno	0,00039
Rosenje	0,00285	Bliskanje	0,00037
PovprecenZracniTlak	0,00271	PovprecnaHitrostVetra	0,00036
DezSSnegom	0,00226	RosenjeKiZmrzuje	0,00036
PovprecnaHitrostVetra	0,00215	MocanVeter	0,00022
DezKiZmrzuje	0,00214	Grmenje	0,00009
Toca	0,00169	Toca	0,00003
Nevihta	0,00135		
BabjePseno	0,00121		
DnevnaKolicinaPadavin	0,00118		
MocanVeter	0,00115		
Sneg	0,00089		
RosenjeKiZmrzuje	0,00070		
MeglaZVidnimNebom	0,00068		
PlohaDezja	0,00067		
Grmenje	0,00047		
Bliskanje	0,00031		
LedeneIglice	0,00031		
Sodra	0,00028		
Dez	0,00018		
ZrnatSneg	0,00004		
LedenaMegla	0,00003		

5.2.3. Pričakovana razlika variance

Za ocenitev atributov pri regresijskem problemu smo uporabili metodo pričakovane razlike varianc. Metoda je primerna za ocenjevanje diskretnih atributov. Vremenske atribute smo zaradi tega uporabili v njihovi logični obliki (true, false). V primeru, da neka lastnost na

opazovani lokaciji prevladuje, ima atribut vrednost »true«, sicer ima atribut vrednost »false«. Pričakovano razliko variance podamo z naslednjo formulo (povzeto po [7]):

$$ds^2(A_i) = \frac{1}{n}\sum_{k=1}^{n}(r^{(k)} - \bar{r})^2 - \sum_{j=1}^{n_i}(p_{.j}\frac{1}{n_{.j}}\sum_{k=1}^{n_{.j}}(r_j^{(k)} - \overline{r_j})^2)$$

n – število učnih primerov

$\bar{r}$– povprečna vrednost zveznega razreda med n učnimi primeri

$r_j^{(k)}$ – vrednost odvisne spremenljivke k-tega primera, ki ima j-to vrednost atributa A_i

$\overline{r_j}$ – povprečna vrednost odvisne spremenljivke primerov z j-to vrednostjo atributa A_i.

Preostali simboli so pojasnjeni v razdelku 5.2.1.

Rezultat ocene atributov z metodo pričakovane razlike variance za nezgode je podan v tabeli 6. Oceno atributov za bolezni najdemo v dodatku A.3. Atributi, ocenjeni kot nepomembni, so v tabelah odstranjeni. Preostali atributi so razvrščeni v padajočem vrstnem redu glede na njihovo pomembnost.

Tabela 6: Ocena atributov s pričakovano razliko variance pri nezgodah.

Število nezgod		**Povprečno izplačilo nezgode**	
atribut	**Ocena**	**atribut**	**ocena**
DanVTednu	0,00069	DanVTednu	0,00113
LetniCas	0,00060	LetniCas	0,00098
LuninaMena	0,00059	LuninaMena	0,00098
Rosa	0,00042	Dez	0,00067
Dez	0,00041	Rosa	0,00067
Padavine	0,00040	Padavine	0,00066
SneznaOdeja	0,00040	SneznaOdeja	0,00066
PlohaDezja	0,00040	Delovnik	0,00066
DezKiZmrzuje	0,00040	Slana	0,00066
Sneg	0,00040	Sneg	0,00065
Slana	0,00040	DezKiZmrzuje	0,00065
Rosenje	0,00040	Nevihta	0,00065
Delovnik	0,00040	PlohaDezja	0,00065
Meglica	0,00040	Meglica	0,00065
Grmenje	0,00039	Ivje	0,00065
Ivje	0,00039	Rosenje	0,00065
Nevihta	0,00039	Grmenje	0,00065
ViharniVeter	0,00039	ViharniVeter	0,00065
TrdoIvje	0,00039	TrdoIvje	0,00065

5.2.4. Regresijski ReliefF

Kot naslednji predstavnik regresijskih evalvatorjev je bil uporabljen algoritem regresijski ReliefF (v nadaljevanju RReliefF). Povzeto po [7] pri regresijskih problemih napovedujemo zvezne vrednosti in zato za napovedovanje ne moremo uporabiti najbližjih pogreškov in zadetkov iz algoritma ReliefF. RReliefF uporablja neke vrste »verjetnost, da dva primera pripadata različnima razredoma«. Algoritem kvaliteto atributa oceni glede na lokalne informacije o razločevanju razredov. Podrobnejšo razlago algoritma najdemo v [16].

Rezultati ocene atributov z algoritmom RRelieF za nezgode so podani v tabeli 7. Tabelo bolezni za algoritem RReliefF najdemo v dodatku A.4. Algoritem smo poganjali toliko časa, dokler je vračal nepomembne atribute. Nepomembne atribute smo ob tem iz vhodne množice atributov sproti odstranjevali. V tabeli so prikazani zgolj atributi, ocenjeni kot pomembni. Ostali atributi so razvrščeni v padajočem vrstnem redu glede na njihovo pomembnost.

Tabela 7: Ocena atributov z RReliefF pri nezgodah.

Število nezgod		**Povprečno izplačilo nezgode**	
atribut	**Ocena**	**atribut**	**ocena**
Sneg	0,00428	PovprecnaHitrostVetra	0,02533
MinimalnaTemperaturaZrakaNa5cm	0,00307	MinimalnaTemperaturaZrakaNa5cm	0,00869
SkupnaVisinaSnezneOdeje	0,00284	LetniCas	0,00304
Grmenje	0,00266	MaksimalnaTemperaturaZrakaNa2m	0,00145
Delovnik	0,00247	TrajanjeSoncnegaObsevanja	0,00143
DnevnaKolicinaPadavin	0,00230	SkupnaVisinaSnezneOdeje	0,00141
MaksimalnaTemperaturaZrakaNa2m	0,00225	PovprecnaTemperaturaZrakaNa2m	0,00139
VisinaNovozapadlegaSnega	0,00223	Delovnik	0,00099
Nevihta	0,00213	DnevnaKolicinaPadavin	0,00099
Dez	0,00212	MocanVeter	0,00095
LuninaMena	0,00207	Teden	0,00074
PovprecnaTemperaturaZrakaNa2m	0,00194	Mesec	0,00070
LetniCas	0,00172	LuninaMena	0,00064
MinimalnaTemperaturaZrakaNa2m	0,00157	PovprecnaOblacnost	0,00061
DezKiZmrzuje	0,00144	PovprecnaRelativnaVlaga	0,00061
PlohaDezja	0,00134	MinimalnaTemperaturaZrakaNa2m	0,00059
TrajanjeSoncnegaObsevanja	0,00112	DanVTednu	0,00047
PovprecnaHitrostVetra	0,00112	PovprecenZracniTlak	0,00045
LedeneIglice	0,00107	VisinaNovozapadlegaSnega	0,00039
PovprecnaOblacnost	0,00095		
PovprecnaRelativnaVlaga	0,00089		
Rosenje	0,00086		
ViharniVeter	0,00064		
Mesec	0,00060		
Teden	0,00059		
MocanVeter	0,00047		
DanVTednu	0,00045		
PovprecenZracniTlak	0,00031		
RosenjeKiZmrzuje	0,00017		
Bliskanje	0,00012		

5.2.5. Rezultati evalvacije

Ocenjevanje atributov lahko pomaga pri poenostavitvi modela. Atribute, ki so se izkazali kot nepomembni, v takšnem primeru iz modela odstranimo. Z odstranitvijo atributov naj ne bi povzročili prevelike škode modelu. Odstranjevanje atributov pride v poštev predvsem v primerih, ko imamo težave zaradi premalo zmogljive strojne opreme. V takšnih primerih je zmanjšanje kompleksnosti modela še kako dobrodošlo.

V našem primeru se je izkazalo, da problem kot tak ni prezahteven, zato iz modela nismo izvzeli nobenega atributa. Po drugi strani ocenjevanje atributov pripomore tudi k boljšemu razumevanju modela.

5.3. Klasifikacija

Odločili smo se, da najprej poskusimo reševati klasifikacijski problem. Klasifikacija je (povzeto po [4]) definirana kot proces iskanja modela, ki opisuje problem in razlikuje med njegovimi razredi. Model je zgrajen z namenom, da je sposoben napovedati umestitev še nerazporejenega objekta v njemu ustrezni razred. Pri klasifikaciji gre za proces v dveh korakih. V prvem koraku se na podlagi učne množice zgradi model klasifikatorja. V drugem koraku gre za klasifikacijo testne množice na podlagi v prvem koraku zgrajenega modela.

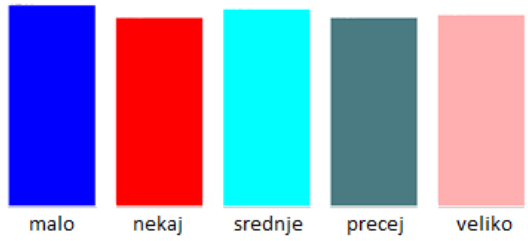
5.3.1. Diskretizacija razreda

Pri klasifikacijskem problemu mora biti razred definiran kot diskretna spremenljivka, kar pomeni, da prvotna oblika podatkov v našem primeru ni bila ustrezna. Razred je bilo najprej potrebno diskretizirati. Glede na Panjana v [12] to pomeni, da je potrebno zvezni interval razreda razdeliti na manjše število podintervalov. Znotraj teh podintervalov so potem vse zvezne vrednosti definirane z isto diskretno vrednostjo. Posledica takšne diskretizacije se kaže tudi v izgubi informacije.

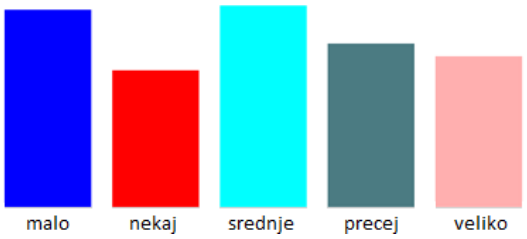
Odločili smo se, da interval razreda razdelimo v pet enakomerno razporejenih podintervalov. V našem primeru skušamo napovedovati dve različni ciljni spremenljivki, zato poglejmo podrobneje vsako posebej.

Pri prvem problemu nam razred predstavlja število prijavljenih odškodninskih zahtevkov. V obzir za klasifikacijo smo vzeli normalizirane vrednosti razreda, gledano od začetka leta 2000 do konca leta 2015. Vrednosti razredov smo uredili po velikosti in množico tako za primer nezgod kot tudi za primer bolezni razrezali na pet številčno karseda enakovrednih skupin.

Zalogo vrednosti posameznega razreda po novem predstavlja pet diskretnih vrednosti. Za primer nezgod je razporeditev predstavljena na sliki 4; za primer bolezni je razporeditev predstavljena na sliki 5.



Slika 4: Zaloga vrednosti števila prijavljenih odškodninskih zahtevkov za nezgode.



Slika 5: Zaloga vrednosti števila prijavljenih odškodninskih zahtevkov za bolezni.

Pri drugem problemu nam razred predstavlja povprečna višina izplačane odškodnine. Tudi v tem primeru vrednosti razreda razbijemo na pet čimbolj enakomerno porazdeljenih skupin. Nove zaloge vrednosti razredov za primer nezgod najdemo na sliki 6, za primer bolezni pa na sliki 7.

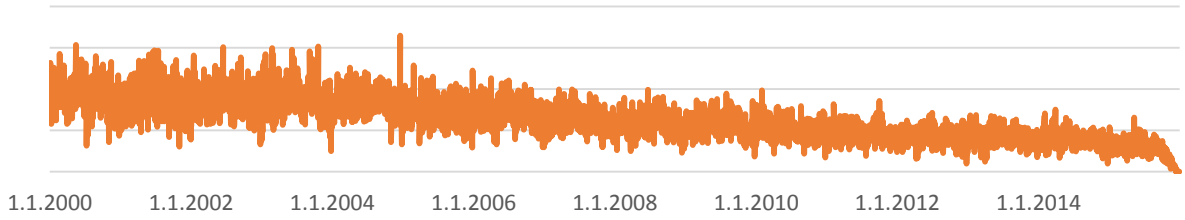
Slika 6: Zaloga vrednosti povprečne višine izplačane odškodnine za nezgode.

Slika 7: Zaloga vrednosti povprečne višina izplačane odškodnine za bolezni.

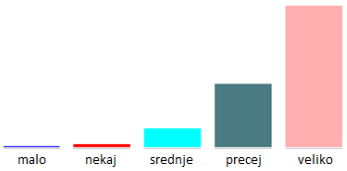
5.3.2. Pregled razporeditve razredov

Ko gledamo podatke od leta 2000 do leta 2015, hitro zbode v oči, da se je porazdelitev razredov skozi leta bistveno spremenila. V prvih letih je bila porazdelitev razredov precej drugačna, kot kaže trenutno stanje.

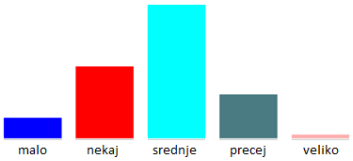
Grafi nezgod v nadaljevanju nazorneje ponazarjajo problem. Graf na sliki 8 prikazuje število prijav nezgod za celotno opazovano obdobje; grafi na slikah 9, 10 in 11 prikazujejo razporeditev razredov za nekaj izbranih let.



Slika 8: Število prijavljenih nezgod od leta 2000 do leta 2015.



Slika 9: Število nezgod v letu 2000.

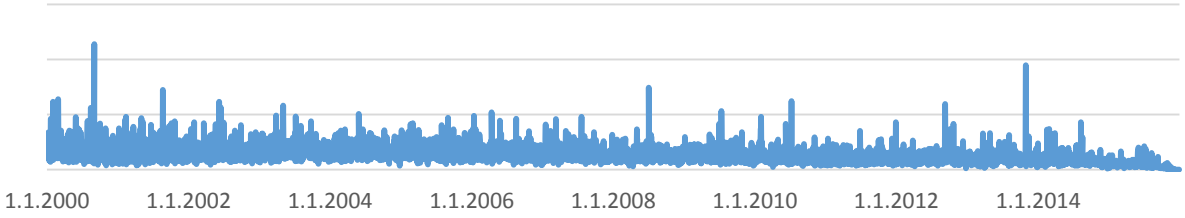


Slika 10: Število nezgod v letu 2008.

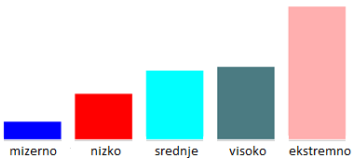
Slika 11: Število nezgod v letu 2015.

Že zgolj iz podanih nekaj let je opaziti, da je celo v kratkem obdobju šestnajstih let trend šel v popolnoma drugo smer. V letu 2000 se je dnevno večinoma prijavljalo veliko nezgod, v letu 2015 pa je število dnevno prijavljenih nezgod bistveno nižje. Tudi na grafih za vmesna obdobja je možno videti preobrat.

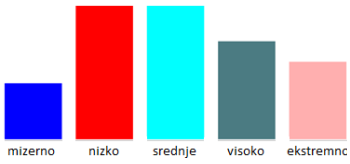
Poglejmo še grafe za povprečno višino izplačane nezgode. Graf na sliki 12 prikazuje višino izplačane nezgode za celotno opazovano obdobje; grafi na slikah 13, 14 in 15 prikazujejo razporeditev razredov za nekaj izbranih let.



Slika 12: Povprečna višina izplačane nezgode od leta 2000 do leta 2015.



Slika 13: Povprečno izplačilo nezgode v letu 2000.



Slika 14: Povprečno izplačilo nezgode v letu 2008.

Slika 15: Povprečno izplačilo nezgode v letu 2015.

V primeru razreda povprečne višine izplačila nezgode je prav tako videti, da se je trend od leta 2000 do leta 2015 obrnil. V letu 2000 je bila večina izplačil nezgod ekstremno visokih, medtem

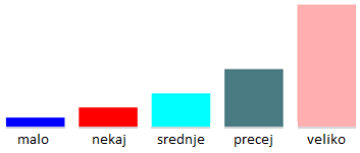
ko so v letu 2015 skoraj samo še mizerna izplačila nezgod. O razlogih za obrnitev trenda več v nadaljevanju.

Vzporedne grafe za vzrok bolezni najdemo v prilogah B.1 in B.2. Tudi pri boleznih je prišlo do obrnitve trenda, vendar se ta obrat zgodi prej in je opazen v bistveno manjši meri kot pri nezgodah.

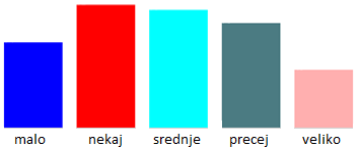
5.3.3. Omejitev obsega podatkov

Ugotovitve iz prejšnjega razdelka nam dajo slutiti, da rezultati, pridobljeni na podlagi takšnih podatkov, ne bodo reprezentativni. Zaradi tolikšnega obrata trenda nima posebnega smisla učno množico graditi na podlagi podatkov od leta 2000 naprej. Sprejeta je bila odločitev, da se analizira podatke šele od vključno leta 2010 naprej, ko je obračanje zaznavno v manjši meri. Porazdelitev razredov je bila prilagojena novim pogojem in je tudi v tem primeru enakomerna.

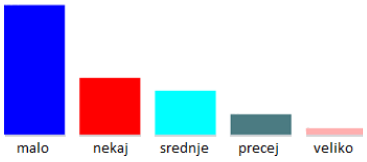
V nadaljevanju podajamo grafe za obdobje od začetka leta 2010 do konca leta 2015. Najprej so podani grafi za število prijavljenih nezgod (slike 16, 17 in 18), nato sledijo grafi za povprečno višino izplačane nezgode (slike 19, 20 in 21).



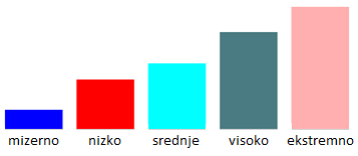
Slika 16: Število nezgod v letu 2010.



Slika 17: Število nezgod v letu 2013.

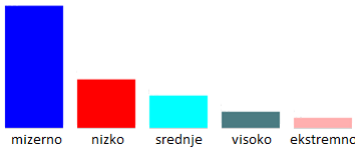


Slika 18: Število nezgod v letu 2015.



Slika 19: Povprečno izplačilo nezgode v letu 2010.

Slika 20: Povprečno izplačilo nezgode v letu 2013.



Slika 21: Povprečno izplačilo nezgode v letu 2015.

Kakor vidimo iz grafov, se nam pri nezgodah tudi pri opazovanju krajšega obdobja pojavi problem obrnitve trendov. Na podlagi tega je bila sprejeta odločitev, da bo pri algoritmih za učno množico vedno upoštevano zgolj obdobje za eno leto nazaj. Na podlagi podatkov za eno leto nazaj od datuma opazovanega problema je zgrajena učna množica in nato izvedena klasifikacija tega problema.

Grafi razporeditev za vzrok bolezni so v prilogah B.3 in B.4. Pri boleznih trenda obračanja pri skrajšanem obdobju ni opaziti.

5.3.4. Klasifikacijski algoritmi

Tehnika klasifikacije je zmožna procesiranja podatkov večjih raznolikosti kakor regresija in zato njena popularnost narašča (povzeto po [14]). Poznamo več vrst metod strojnega učenja. Delijo se glede na to, kaj je njihov rezultat učenja. Metode, skupaj z algoritmi, uporabljenimi v tem magistrskem delu, navajamo v nadaljevanju. Navajamo zgolj povzetke uporabljenih metod klasifikacije; podrobnejše opise najde bralec v [7]. Dober povzetek metod najdemo tudi v delu Štrausa [19].

Bayesov klasifikator

Bayesov klasifikator deluje na principu izračuna pogojne verjetnosti za vsak razred pri danih vrednostih (vseh) atributov za problem.

Wekin algoritem, uporabljen iz te metode, se imenuje »NaiveBayes«. Značilnost algoritma je, da predpostavi pogojno neodvisnost atributov.

Odločitvena drevesa

Odločitvena drevesa so dobila ime po diagramu v obliki drevesa. Sodijo v skupino metod nadzorovanega avtomatskega učenja, pri katerih učenje deluje po metodi deli in vladaj.

Uporabljeni algoritem iz te metode se imenuje »J48«. Gre za odprtokodno implementacijo algoritma C4.5 iz Weke.

Ansambelske metode

Osnovna ideja te metode je združevanje klasifikatorjev. Zgradimo več klasifikatorjev, ki jim dovolimo glasovati za končno odločitev.

Med ansambelskimi metodami smo izbrali Breimanov algoritem Naključni gozdovi [2]. V Weki je algoritem poznan pod imenom »RandomForest«. Izhaja iz ideje bagginga [2], ki je prav tako Breimanovo delo. Glede na [7] pri baggingu za učno množico z n primeri n krat naključno izberemo primer iz učne množice in v okviru vsake take množice potem zaženemo učni algoritem. Naključni gozdovi veljajo za nadgradnjo bagginga (opis povzet po [24]). Iz učne množice naredimo izbrano število novih učnih množic, v okviru katerih zgradimo odločitvena drevesa. Atribut za razcep vozlišča je vsakič izmed naključno izbranimi atributi; napovedi več dreves združimo glede na večino.

Najbližji sosedi

Predstavlja eno izmed najstarejših metod. Metoda iz učne množice poišče nekaj primerov, ki so novemu primeru najbolj podobni (najbližji).

Iz te skupine metod smo uporabili algoritem IBk. Gre za predstavnika algoritmov k najbližjih sosedov (ang. K-nearest neighbor), ki je na voljo v orodju Weka.

Opisane algoritme smo uporabili pri reševanju klasifikacijskega problema. Testna množica je zajemala podatke od 1. 11. 2011 do 31. 12. 2015. Učna množica je bila za vsak testni primer generirana posebej; vanjo so bili zajeti podatki za eno leto nazaj. Tako je bila na primer za testni primer na datum 7. 3. 2014 generirana učna množica, v katero so bili zajeti podatki od 7. 3. 2013 do 6. 3. 2014. V tabeli 8 so podani rezultati primerjave različnih klasifikacijskih algoritmov.

Tabela 8: Rezultati klasifikacije za različne algoritme.

vzrok	razred	večinski razred	NaiveBayes	J48	RandomForest	IBk
NEZGODE	Stevilo	20,26%	30,41%	30,47%	36,88%	33,92%
NEZGODE	Izplacilo	20,77%	27,73%	27,29%	31,29%	32,44%
BOLEZNI	Stevilo	22,41%	21,48%	20,93%	24,38%	26,30%
BOLEZNI	Izplacilo	20,90%	22,03%	22,03%	20,93%	21,53%

Poleg odstotka pravilno klasificiranih instanc različnih algoritmov so v tabeli podani tudi odstotki večinskega razreda učnih množic. V primerih, ko odstotek pravilno klasificiranih instanc ne presega odstotka večinskega razreda, smo polja obarvali rdeče. V takšnih primerih je mogoče govoriti o neuspešnosti klasifikacijskega algoritma. Na splošno sta se najbolje odrezala algoritma RandomForest in IBk. Algoritem RandomForest celo najbolje, če gledamo samo oceno napovedi za nezgode. Kljub vsemu ugotavljamo, da so odstotki pravilno klasificiranih instanc prenizki in da nas postopek klasifikacije ni pripeljal do želenega rezultata.

Kononeko navaja [7], da klasifikacijska točnost ni najboljša mera napake. Problem je v tem, da klasifikacijska točnost ničesar ne pove o tem, kako dobro so klasificirani primeri iz posameznih razredov. Ker ni upoštevana razdalja med razredi, ne vemo, kako blizu meje razreda je bil ocenjen določeni primer. Odločili smo se, da se preskusimo še v reševanju regresijskega problema, saj so vsi štirje problemi po naravi regresijski.

5.4. Regresija

Pri regresiji za razliko od klasifikacije ne operiramo več z diskretnimi razredi, ampak za ciljni atribut uporabimo zvezne vrednosti. Tu gre za napovedovanje konkretnih vrednosti ciljnega atributa.

Pri klasifikaciji se del informacije izgubi. Možno da je klasifikacija instanco umestila v napačen razred, kljub temu da je bila ocenjena blizu območju pravilnega razreda. Če je pri klasifikaciji velika večina instanc ocenjena blizu pravilne napovedi razreda, je kljub vsemu možno govoriti o dobrem delovanju algoritma za napovedovanje. Regresija nam razjasni takšne nejasnosti, saj napoveduje dejanske vrednosti.

5.4.1. Regresijski algoritmi

Že v razdelku 5.3.4 smo pisali o metodah strojnega učenja. Tam so bile omenjene metode, uporabljene za reševanje klasifikacijskega problema. Na tem mestu opišimo še metode, ki smo jih uporabili za reševanje regresijskega problema. Poleg metod navajamo tudi algoritme, ki smo jih uporabili.

Zavoljo možnosti primerjave med reševanjem klasifikacijskih in regresijskih problemov smo pri reševanju regresijskega problema uporabili algoritme podobnih metod, kot smo jih uporabili pri klasifikaciji. Podrobnejše opise metod najdemo v [7] in [19].

Regresijska drevesa

Metoda regresijskih dreves je analogna klasifikacijski metodi odločitvenih dreves, ki smo jo opisali v razdelku 5.3.4.

Za reševanje regresijskega problema smo uporabili Wekin algoritem »REPTree«. REPTree je hiter algoritem regresijskih dreves, ki drevo gradi na podlagi razlike variance. Algoritem je dobil ime po metodi, ki jo uporablja za obrezovanje drevesa (Reduced Error Pruning Tree) [5].

Ansambelske metode

Analogna klasifikacijska metoda je opisana v razdelku 5.3.4.

Kot predstavnika te metode smo uporabili Wekin algoritem »RandomForest«. Analogen algoritem je bil uporabljen tudi pri klasifikaciji, saj RandomForest podpira reševanje tako regresijskih kot tudi klasifikacijskih problemov.

Najbližji sosedi

Analogna klasifikacijska metoda je opisana v razdelku 5.3.4.

Tudi pri regresiji smo uporabili že iz klasifikacije poznan Wekin algoritem »IBk«. Algoritem IBk je namenjen reševanju regresijskih in klasifikacijskih problemov.

Linearna regresija

Algoritem linearne regresije je analogen algoritmu NaiveBayes pri klasifikaciji. Linearna regresija je najbolj uporabljana med vsemi regresijskimi metodami. Deluje na principu modeliranja odnosa med odvisnimi in neodvisnimi spremenljivkami [12], tako da zgradimo linearen model. Slabost metode je, da ne opisuje najbolje odvisnosti, ki ni linearna.

Uporabljeni Wekin algoritem je znan pod imenom »LinearRegression«.

Navedene algoritme smo uporabili za napovedovanje regresijskega problema. Kot že pri klasifikaciji smo tudi tukaj upoštevali le podatke od začetka leta 2010 do konca leta 2015. V tabeli 9 najdemo ocene regresijskega modela; razlaga tabele je podana v nadaljevanju.

Tabela 9: Rezultati regresije za različne algoritme.

vzrok	cilj	algoritem	RMAE	RRMSE
NEZGODE	Stevilo	LinearRegression	0,7634	**0,7830**
NEZGODE	Stevilo	REPTree	0,8298	**0,8396**
NEZGODE	Stevilo	RandomForest	0,7288	**0,7410**
NEZGODE	Stevilo	IBk	0,7655	**0,7733**
NEZGODE	Izplacilo	LinearRegression	0,9479	**0,9671**
NEZGODE	Izplacilo	REPTree	0,9478	**0,9651**
NEZGODE	Izplacilo	RandomForest	0,8974	**0,9189**
NEZGODE	Izplacilo	IBk	0,8943	**0,9357**
BOLEZNI	Stevilo	LinearRegression	1,0659	**1,0816**
BOLEZNI	Stevilo	REPTree	1,0286	**1,0337**
BOLEZNI	Stevilo	RandomForest	1,0233	**1,0206**
BOLEZNI	Stevilo	IBk	1,0283	**1,0297**
BOLEZNI	Izplacilo	LinearRegression	1,1265	**1,0560**
BOLEZNI	Izplacilo	REPTree	1,0650	**1,2715**
BOLEZNI	Izplacilo	RandomForest	1,0698	**1,0311**
BOLEZNI	Izplacilo	IBk	1,0670	**1,0443**

Za ocenitev stopnje napake smo se pri napovedih uporabili dve izmed pogosto uporabljanih meritev za ocenjevanje učenja regresijskih problemov. Prva meritev se imenuje relativna srednja absolutna napaka (angl. Relative mean absolute error) in jo predstavimo z oznako RMAE [7]. Relativna srednja absolutna napaka je nenegativna in je manjša od 1,0 za sprejemljive hipoteze.

Formula za izračun RMAE je naslednja:

$$RMAE = \frac{\sum_{i=1}^{n}|f(i) - \hat{f}(i)|}{\sum_{i=1}^{n}|f(i) - \bar{f}|}$$

n – število testnih primerov

$f(i)$ – dejanska vrednost

$\hat{f}(i)$ – napovedana vrednost

$\bar{f}$ – povprečna vrednost $\hat{f}(i)$

Druga uporabljena regresijska meritev se imenuje koren relativne srednje kvadratne napake (angl. Relative root mean squared error) in jo predstavimo z oznako RRMSE [8]. Tudi za relativno srednjo kvadratno napako velja, da je nenegativna in manjša od 1,0 za sprejemljive hipoteze. Formula za izračun RRMSE je naslednja:

$$RRMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(f(i) - \hat{f}(i)\right)^2}{\sum_{i=1}^{n}\left(f(i) - \bar{f}\right)^2}}$$

Uporabljeni simboli so enaki kot pri predhodno opisani formuli RMAE.

Za obe formuli je značilno, da bližje, kot se vrednost ocene približa 0,0, bolje se algoritem obnese. Če vrednost ocene preseže 1,0, pomeni da tak algoritem ni učinkovit in da je napoved slabša od napovedi povprečne vrednosti. To pomeni, da bo napoved s tako ocenjenim algoritmom neuporabna in da takšna hipoteza ni sprejemljiva. Več o metodah meritev za napovedovanje najdemo v članku Prestwicha in ostalih [15].

Pri pregledu tabele 9 ugotovimo, da se je najbolje odrezal algoritem RandomForest. Kljub vsemu so vrednosti meritev previsoke in napovedi zato niso zanesljive. Pri meritvah bolezni so celo vse vrednosti nad vrednostjo 1,0, kar pomeni da algoritem za napovedovanje tega problema sploh ni uporaben. O razlogih za to več v nadaljevanju.

Rezultati regresije so potrdili rezultate klasifikacije, ki so se prav tako izkazali za ne preveč uporabne. Algoritem RandomForest se je v tej oceni v obeh primerih v povprečju odrezal najbolje. Pri nadaljnjem delu smo v tem magistrskem delu zato uporabljali zgolj še algoritem RandomForest.

5.5. Časovno okno

Do sedaj smo v algoritmih kot časovno okno podatkov za učno množico uporabljali pogled nazaj za eno leto. V razdelku 5.3.3 smo na podlagi razporejenosti razredov skozi pretekla leta prišli do sklepa o smiselnosti uporabe takšnega časovnega okna. Raznolikost podatkov med leti je napeljevala k razmišljanju o neprimernosti podatkov, zajetih v večje časovno okno. Da bi si izbrali manjše časovno okno, pa se tudi nekako ni zdelo smiselno, saj pri tem podatek izgubi na informaciji. Na primer, pri polletnem časovnem oknu izgubimo informacijo o vsaj enem letnem času.

Kljub vsemu preverimo, če vseeno morebiti ne bi bilo kakšno drugačno časovno okno ustreznejše. Zato smo ponovili postopek regresije z algoritmom RandomForest ob uporabi različnih časovnih oken. Preizkusili smo razširjanje časovnega okna in tudi njegovo krčenje. Na osnovi tega smo dobili potrditev, ali so naša predvidevanja o letnem časovnem oknu pravilna. Za časovno okno smo upoštevali okna naslednjih velikosti: vse, 2 leti, 1 leto, 6 mesecev, 3 mesece in 1 mesec. Pri časovnem oknu »vse« učna množica obsega vse instance pred datumom, za katerega napovedujemo. Pri ostalih oknih so v učno množico zaobsežene le instance pred tem datumom v velikosti časovnega okna. V tabeli 10 podajamo primerjavo ocene uspešnosti pri različnih časovnih oknih.

Tabela 10: Primerjava uspešnosti algoritma RandomForest pri različnih časovnih oknih.

vzrok	cilj	časovno okno	RMAE	RRMSE
NEZGODE	Stevilo	vse	1,0132	**0,9725**
NEZGODE	Stevilo	2 leti	0,8019	**0,8043**
NEZGODE	Stevilo	1 leto	0,7288	**0,7410**
NEZGODE	Stevilo	6 mesecev	0,6932	**0,7184**
NEZGODE	Stevilo	3 meseci	0,6993	**0,7190**
NEZGODE	Stevilo	1 mesec	0,7038	**0,7273**
NEZGODE	Izplacilo	vse	1,0593	**0,9843**
NEZGODE	Izplacilo	2 leti	0,9812	**0,9440**
NEZGODE	Izplacilo	1 leto	0,8974	**0,9189**
NEZGODE	Izplacilo	6 mesecev	0,8891	**0,9298**
NEZGODE	Izplacilo	3 meseci	0,8934	**0,9363**
NEZGODE	Izplacilo	1 mesec	0,9144	**0,9546**
BOLEZNI	Stevilo	vse	1,0562	**1,0418**
BOLEZNI	Stevilo	2 leti	1,0213	**1,0233**
BOLEZNI	Stevilo	1 leto	1,0233	**1,0206**
BOLEZNI	Stevilo	6 mesecev	1,0194	**1,0227**
BOLEZNI	Stevilo	3 meseci	1,0173	**1,0185**
BOLEZNI	Stevilo	1 mesec	1,0321	**1,0300**
BOLEZNI	Izplacilo	vse	1,0434	**1,0208**
BOLEZNI	Izplacilo	2 leti	1,0626	**1,0305**
BOLEZNI	Izplacilo	1 leto	1,0698	**1,0311**
BOLEZNI	Izplacilo	6 mesecev	1,1520	**1,0871**
BOLEZNI	Izplacilo	3 meseci	1,1501	**1,0756**
BOLEZNI	Izplacilo	1 mesec	1,1540	**1,0939**

Pri pregledu rezultatov ugotovimo, da se je pri prvem problemu najbolje odrezalo časovno okno, ki zajema podatke za šest mesecev nazaj. Odstopanje pri časovnem oknu za eno leto je v tem primeru sicer zaznati, vendar sta rezultata kljub vsemu primerljiva. Pri drugem problemu se je najbolje odrezalo časovno okno, ki zajema prav podatke za eno leto nazaj. Za tretji in četrti problem nismo dobili pri nobeni velikosti okna uporabnih rezultatov, saj so vse vrednosti višje od 1,0.

Naša predvidevanja, da za učno množico vzamemo podatke za eno leto nazaj, niti niso bila tako zmotna. Primerjava različnih časovnih oken to hipotezo potrjuje. Pri posameznih problemih res prihaja do manjših odmikov, vendar - gledano na problem v celoti - se kot časovno okno res najbolj splača uporabiti podatke za eno leto nazaj.

5.6. Časovna skala

Obljubili smo, da bomo uspešnost algoritma preučili tudi za različne časovne skale. Kot enoto instance je mogoče namesto dneva zaobjeti kakšen drugačen interval. Za primerjavo smo upoštevali še dve izmed možnih časovnih skal, in sicer teden in mesec.

Pri spremembi enote časovne skale je potrebna tudi prilagoditev podatkov. Pri večji časovni skali se zajetje večjega intervala odrazi v zmanjšanju količine podatkov. Vremenske atribute je potrebno povprečiti, da odražajo povprečne vrednosti vremena za opazovano obdobje izbrane časovne skale. Diskretne atribute predstavimo s prevladujočo vrednostjo v opazovanem obdobju. Nekatere atribute moramo v primeru večanja časovne skale zavreči. Atribut lunina mena, denimo, še ima nek smisel pri tedenski skali, saj ga je mogoče predstaviti s prevladujočo vrednostjo v tednu. Pri mesečni časovni skali atribut lunina mena ne poda več nobene informacije, saj se v obdobju enega meseca izmenjajo vse faze lunine mene. Tudi podatek o delovnih dneh se v primeru mesečne časovne skale izkaže kot neuporaben. Atribut v takih primerih zavržemo. V tabeli 11 podajamo primerjavo uspešnosti algoritma za različne časovne skale.

Tabela 11: Primerjava ocene uspešnosti pri različnih časovnih skalah.

vzrok	cilj	časovna skala	RMAE	RRMSE
NEZGODE	Stevilo	dan	0,7288	**0,7410**
NEZGODE	Stevilo	teden	0,6142	**0,6314**
NEZGODE	Stevilo	mesec	0,5998	**0,6177**
NEZGODE	Izplacilo	dan	0,8974	**0,9189**
NEZGODE	Izplacilo	teden	0,7915	**0,7836**
NEZGODE	Izplacilo	mesec	0,7017	**0,6830**
BOLEZNI	Stevilo	dan	1,0233	**1,0206**
BOLEZNI	Stevilo	teden	1,0058	**1,0058**
BOLEZNI	Stevilo	mesec	0,9258	**1,0406**
BOLEZNI	Izplacilo	dan	1,0698	**1,0311**
BOLEZNI	Izplacilo	teden	1,0943	**1,0569**
BOLEZNI	Izplacilo	mesec	1,0272	**1,0465**

Pri pregledu rezultatov opazimo, da se je mesečna časovna skala pri nezgodah izkazala za najuspešnejšo. Tudi pri boleznih se je v enem primeru mesec pojavil kot edina sprejemljiva možnost; pri ostalih ocenah smo dobili vrednosti višje od 1,0 in so zato rezultati neuporabni. To napeljuje k razmišljanju, da naj dnevna časovna skala ne bi bila najbolje izbrana. Vendar, upoštevati je potrebno še drugo plat. Vprašanje je, ali so za nas napovedi na nivoju meseca

sploh sprejemljive. Če da, potem bi to pomenilo večjo uspešnost pri napovedovanju. Razumljivo je, da bolj kot podatke generaliziramo, preprosteje jih je napovedovati. Vendar mi skušamo prikazati vpliv vremena na odškodninske zahtevke, zato je vseeno bolje ohraniti dnevno časovno skalo. Le pri dnevni časovni skali je možno preučevati vse podnebne spremenljivke v odnosu do odškodninskih zahtevkov.

5.7. Lokalizacija

Kot eden izmed glavnih prispevkov magistrskega dela je bila v uvodnem delu omenjena zmožnost lokalnega napovedovanja odškodninskih zahtevkov. Na tem mestu skušamo predstaviti različne pristope k poskusu te lokalizacije. Prikazano je več poskusov razreza Slovenije na manjše enote; poleg tega je podana ocena uspešnosti vsakega razreza.

Žal so se vsi poskusi lokalizacije izkazali za bolj ali manj neuspešne. O razlogih za to več v nadaljevanju.

5.7.1. Statistične regije

Kot prva možnost razreza Slovenije na manjše enote se ponuja razrez po statističnih regijah. Glede na Statistični urad republike Slovenije (v nadaljevanju SURS) [29] imamo pri nas 12 statističnih regij:

- Pomurska,
- Podravska,
- Koroška,
- Savinjska,
- Zasavska,
- Spodnjeposavska,
- Jugovzhodna Slovenija,
- Primorsko-kraška,
- Osrednjeslovenska,
- Gorenjska,
- Goriška,
- Obalno-kraška.

Na sliki 22 je prikazan razrez Slovenije glede na statistične regije. Za vsako občino je natančno določeno, kateri statistični regiji pripada. Umestitev kraja v ustrezno statistično regijo je bilo izvedeno glede na pripadnost občini.

Slika 22: Porazdelitev Slovenije glede na statistične regije (Vir: http://www.slora.si/definicije-kazalnikov-in-metod).

V tabeli 12 podajamo rezultat ocene algoritma RandomForest pri uporabljenem razrezu. Narejena je primerjava med dano oceno algoritma za celotno Slovenijo in ocenami posameznih razrezov po statističnih regijah Slovenije. Opaziti je, da se pri lokalizaciji rezultat ocene uspešnosti poslabša. Vrednosti napake RRMSE so v večini primerov nad vrednostjo 1,0, kar pomeni, da ta način razreza označimo kot neuporaben.

Tabela 12: Uspešnost algoritma RandomForest pri razrezu na statistične regije.

	NEZGODE Stevilo		NEZGODE Izplacilo		BOLEZNI Stevilo		BOLEZNI Izplacilo	
REGIJA	**RMAE**	**RRMSE**	**RMAE**	**RRMSE**	**RMAE**	**RRMSE**	**RMAE**	**RRMSE**
SLOVENIJA	0,7288	**0,7410**	0,8974	**0,9189**	1,0233	**1,0206**	1,0698	**1,0311**
GORENJSKA	0,9455	**0,9411**	1,0490	**1,0133**	1,0255	**1,0436**	1,0843	**1,0393**
GORIŠKA	0,9919	**0,9733**	1,0668	**1,0231**	1,0190	**1,0354**	1,0796	**1,0475**
JUGOVZHODNA SLOVENIJA	0,9912	**0,9851**	1,0611	**1,0149**	1,0437	**1,0442**	1,0951	**1,0345**
KOROŠKA	1,0161	**1,0226**	1,1198	**1,0373**	1,0670	**1,0394**	1,0828	**1,0295**
OBALNO-KRAŠKA	1,0296	**0,9926**	1,1492	**1,0250**	1,0364	**1,0350**	1,1623	**1,0334**
OSREDNJESLOVENSKA	0,9507	**0,9467**	1,0247	**1,0036**	1,0266	**1,0283**	1,1348	**1,0555**
PODRAVSKA	1,0288	**0,9899**	1,0745	**1,0221**	1,0047	**1,0291**	1,0856	**1,0316**
POMURSKA	0,9840	**0,9606**	1,0801	**1,0437**	1,0098	**1,0293**	1,0662	**1,0376**
SPODNJEPOSAVSKA	1,0355	**0,9954**	1,0453	**1,0171**	1,0413	**1,0387**	1,0679	**1,0182**
PRIMORSKO-KRAŠKA	1,0052	**1,0080**	1,1318	**1,0578**	1,0681	**1,0397**	1,0260	**1,0254**
SAVINJSKA	0,9186	**0,9145**	1,0068	**1,0018**	1,0039	**1,0297**	1,0771	**1,0611**
ZASAVSKA	1,0747	**1,0276**	1,1302	**1,0336**	1,0335	**1,0340**	1,0461	**1,0275**

5.7.2. Strani neba – poševni razrez

Naslednji od možnih načinov razreza Slovenije izvedemo glede na strani neba. Kot središčno točko razreza določimo geometrično središče Slovenije (v nadaljevanju GEOSS), ki je v kraju Vače. GEOSS predstavlja težišče Slovenije in s tem nekako pooseblja središče naše države. Središča sicer nepravilnemu liku, kakršen je obris meja Slovenije, ni mogoče popolnoma natančno določiti.

Pri razrezu skozi GEOSS potegnemo črto pod kotom 45 stopinj in na njo pravokotno potegnemo drugo črto. Na ta način Slovenijo razdelimo na štiri dele glede na strani neba sever, jug, vzhod in zahod (glej sliko 23). Pri umestitvi krajev v ustrezni kvadrant si pomagamo z enačbami razmejitvenih premic. Glede na to, v kateri kvadrant je razporejen opazovani kraj, ga umestimo v ustrezno regijo.

Slika 23: Razrez Slovenije glede na strani neba (S, J, V, Z).

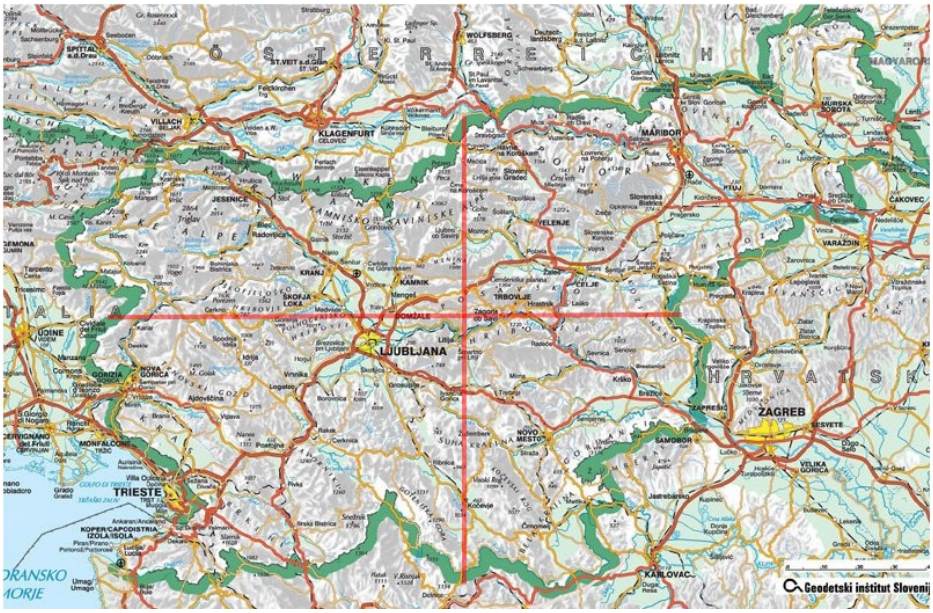
Rezultate algoritma RandomForest pri razrezu glede na strani neba si lahko ogledamo v tabeli 13. Uspešnost algoritma pri razrezu spet kaže poslabšanje rezultata v primerjavi z uspešnostjo za celotno Slovenijo. Metoda se izkaže kot zgolj pogojno uporabna, saj je večina vrednosti RRMSE nad vrednostjo 1,0. Pogojno uporaben bi bil algoritem zgolj za napovedovanje števila nezgod, medtem ko se je algoritem pri napovedovanju bolezni izkazal za popolnoma neuporabnega.

Tabela 13: Uspešnost algoritma RandomForest pri razrezu na strani neba (S, J, V, Z).

	NEZGODE Stevilo		NEZGODE Izplacilo		BOLEZNI Stevilo		BOLEZNI Izplacilo	
REGIJA	RMAE	RRMSE	RMAE	RRMSE	RMAE	RRMSE	RMAE	RRMSE
SLOVENIJA	0,7288	**0,7410**	0,8974	**0,9189**	1,0233	**1,0206**	1,0698	**1,0311**
SEVER	0,9735	**0,9853**	1,0707	**1,0461**	1,0253	**1,0323**	1,0597	**1,0363**
VZHOD	0,8444	**0,8494**	0,9575	**0,9633**	1,0611	**1,0232**	1,0535	**1,0242**
JUG	0,9598	**0,9634**	1,0754	**1,0363**	1,0221	**1,0393**	1,0819	**1,0354**
ZAHOD	0,8483	**0,8375**	0,9883	**0,9893**	1,0246	**1,0313**	1,1241	**1,0704**

5.7.3. Strani neba – prečni razrez

Še ena varianta razreza Slovenije glede na strani neba je razrez na severovzhod, jugovzhod, jugozahod in severozahod. Pri tem razrezu skozi GEOSS potegnemo navpično in vodoravno črto ter tako razdelimo Slovenijo na štiri dele (glej sliko 24). Enako kot pri predhodni razdelitvi glede na strani neba si tudi tukaj pri umestitvi krajev pomagamo z enačbami začrtanih premic. Kraje razmestimo v regije glede na njihovo umestitev v kvadrant.



Slika 24: Razrez Slovenije glede na strani neba (SV, JV, JZ, SZ).

Rezultat razreza za ta primer je razviden v tabeli 14. V primerjavi z rezultatom uspešnosti za celotno Slovenijo smo pri lokalizaciji spet dobili slabše rezultate. Rezultati za nezgode so primerljivi s prejšnjim razrezom, pri boleznih smo ponovno dobili vse vrednosti RRMSE nad vrednostjo 1,0.

Tabela 14: Uspešnost algoritma RandomForest pri razrezu na strani neba (SV, JV, JZ, SZ).

	NEZGODE Stevilo		NEZGODE Izplacilo		BOLEZNI Stevilo		BOLEZNI Izplacilo	
REGIJA	RMAE	RRMSE	RMAE	RRMSE	RMAE	RRMSE	RMAE	RRMSE
SLOVENIJA	0,7288	**0,7410**	0,8974	**0,9189**	1,0233	**1,0206**	1,0698	**1,0311**
SEVEROVZHOD	0,8427	**0,8426**	0,9683	**0,9734**	1,0505	**1,0188**	1,0526	**1,0348**
JUGOVZHOD	0,9505	**0,9544**	1,0281	**0,9985**	1,0170	**1,0423**	1,0779	**1,0235**
JUGOZAHOD	0,8977	**0,8949**	1,0316	**1,0055**	1,0692	**1,0308**	1,1137	**1,0435**
SEVEROZAHOD	0,9159	**0,9152**	1,0168	**0,9882**	1,0692	**1,0402**	1,1021	**1,0405**

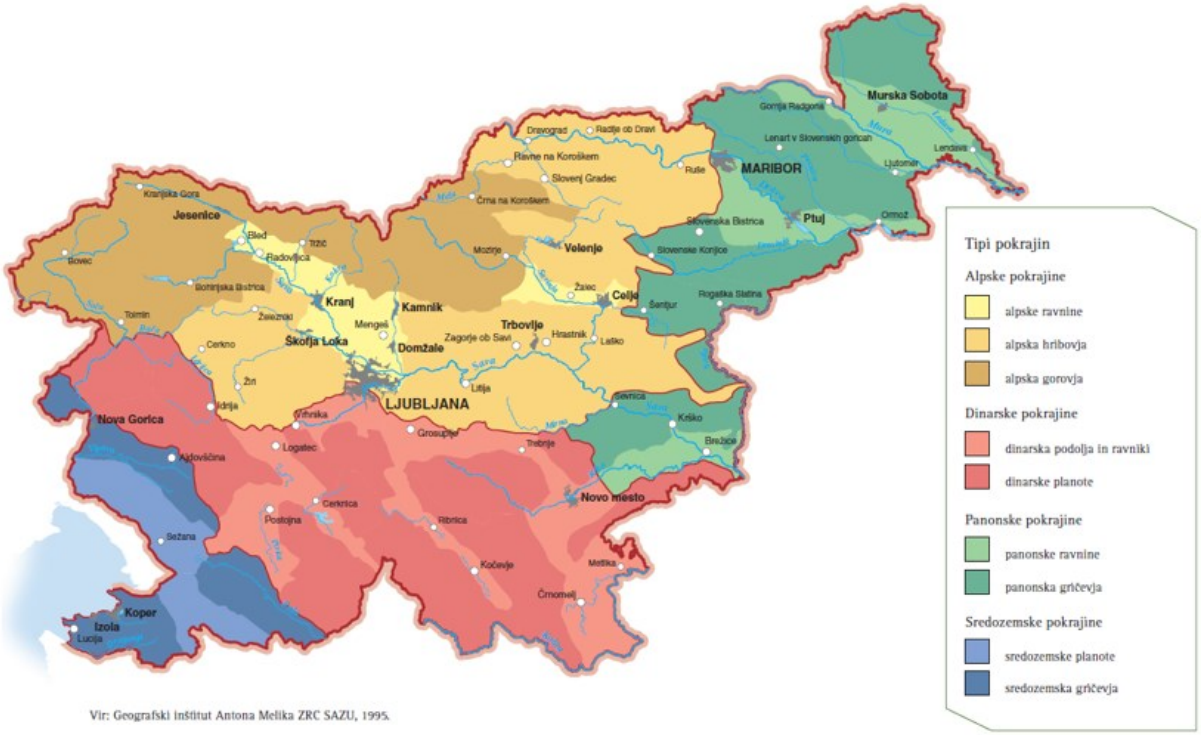
5.7.4. Tip pokrajine

Slovenijo je mogoče razdeliti tudi glede na tipe pokrajin. Glede na vremenski portal [26] poznamo v Sloveniji štiri glavne tipe pokrajin. Tipi se sicer še podrobneje delijo, vendar se v našem primeru osredotočamo zgolj na osnovno razdelitev na štiri osnovne tipe. Osnovna razdelitev je uporabljena zgolj zaradi majhnosti vzorcev pri večji razdrobljenosti. Tipi pokrajin, ki jih bomo obravnavali, so tako naslednji:

- alpska pokrajina,
- dinarska pokrajina,

- panonska pokrajina,
- sredozemska pokrajina.

Razdelitev je izvedena glede na to, v kateri tip pokrajine sodi večinski del opazovane občine, kateri pripada opazovani kraj (glej sliko 25). Nadejali smo se podobnosti vremenskih razmer znotraj pokrajin.



Slika 25: Razdelitev Slovenije glede na tipe pokrajin (Vir: »Javne informacije Slovenije«, ARSO-met [26]).

V tabeli 15 vidimo rezultate pri razdelitvi glede na tipe pokrajin. Opaziti je, da tudi pri tem razrezu ni prišlo do izboljšav glede na rezultate za celotno Slovenijo. Rezultati razreza so podobni prejšnjim razrezom: pogojno uporabni za napovedovanje nezgod in neuporabni za napovedovanje bolezni.

Tabela 15: Uspešnost algoritma RandomForest pri razrezu glede na tip pokrajine.

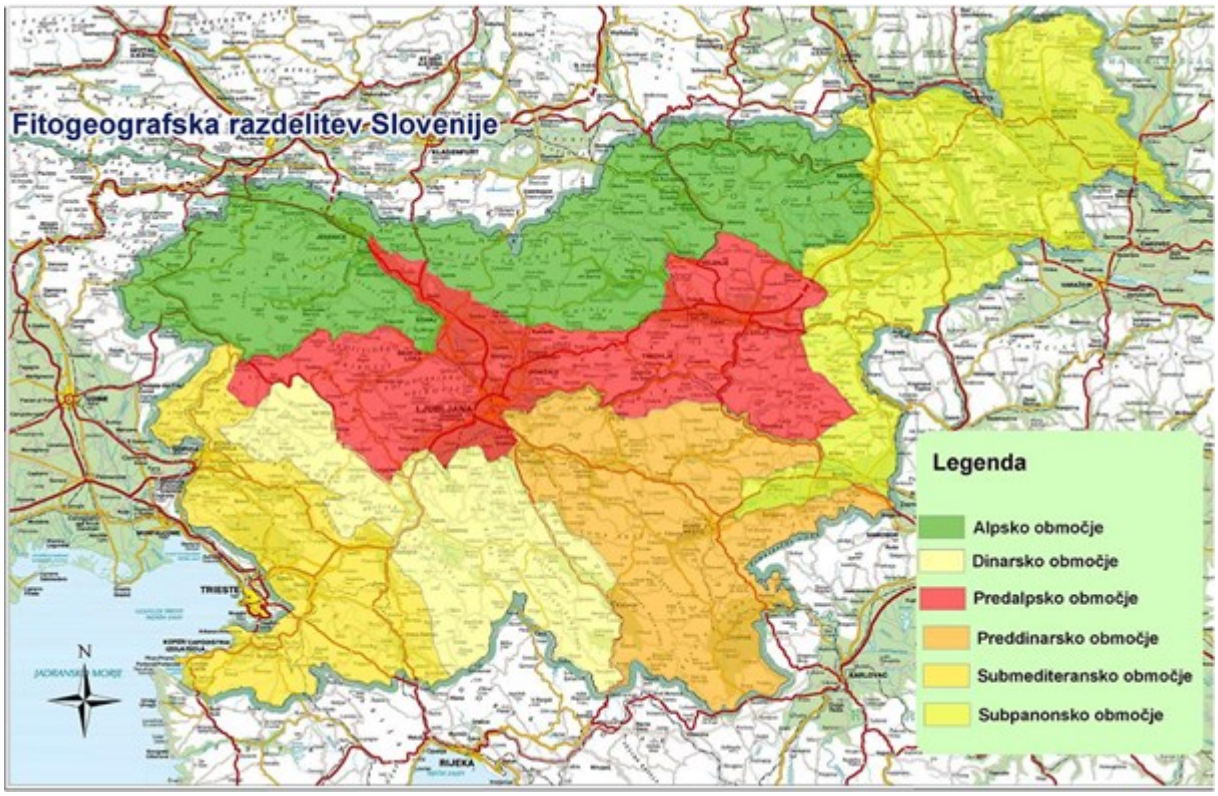
	NEZGODE Stevilo		NEZGODE Izplacilo		BOLEZNI Stevilo		BOLEZNI Izplacilo	
REGIJA	RMAE	RRMSE	RMAE	RRMSE	RMAE	RRMSE	RMAE	RRMSE
SLOVENIJA	0,7288	**0,7410**	0,8974	**0,9189**	1,0233	**1,0206**	1,0698	**1,0311**
PANONSKA	0,8875	**0,8878**	0,9740	**0,9666**	1,0083	**1,0208**	1,0527	**1,0173**
ALPSKA	0,8424	**0,8396**	0,9683	**0,9741**	1,0385	**1,0194**	1,1030	**1,0447**
DINARSKA	0,9324	**0,9388**	1,0465	**1,0075**	1,0302	**1,0461**	1,1457	**1,0547**
SREDOZEMSKA	0,9750	**0,9562**	1,0906	**1,0178**	1,0028	**1,0357**	1,1505	**1,0493**

5.7.5. Fitogeografska območja

Razdelitev Slovenije na fitogeografska območja sta že 1997 opisala Zupančič in Smole [23]. Če povzamemo, je fitogeografska porazdelitev definirana na podlagi klime, geološke podlage z reliefom, vodovja in flore. Poznamo šest območij fitogeografske delitve Slovenije:

- alpsko območje,
- dinarsko območje,
- predalpsko območje,
- preddinarsko območje,
- submediteransko območje,
- subpanonsko območje.

Pri razdelitvi na fitogeografska območja sta upoštevana klima in relief (glej sliko 26). Pričakovali smo, da imajo območja podobne vremenske razmere. Razdelitev je bila izvedena na podlagi ocene, v katero območje spada večinski del opazovane občine, v kateri leži opazovani kraj.



Slika 26: Fitogeografska razdelitev Slovenije (Vir: Uprava RS za varno hrano, veterinarstvo in varstvo rastlin).

Rezultati fitogeografskega razreza so podani v tabeli 16. Kljub podobnosti klime in reliefa je tudi v tem primeru zaslediti poslabšanje rezultata v primerjavi z rezultatom za celotno

Slovenijo. RRMSE v večini primerov presega vrednost 1,0, tako da se je razdelitev pokazala kot neuporabna.

Tabela 16: Uspešnost algoritma RandomForest pri razrezu na fitogeografske regije.

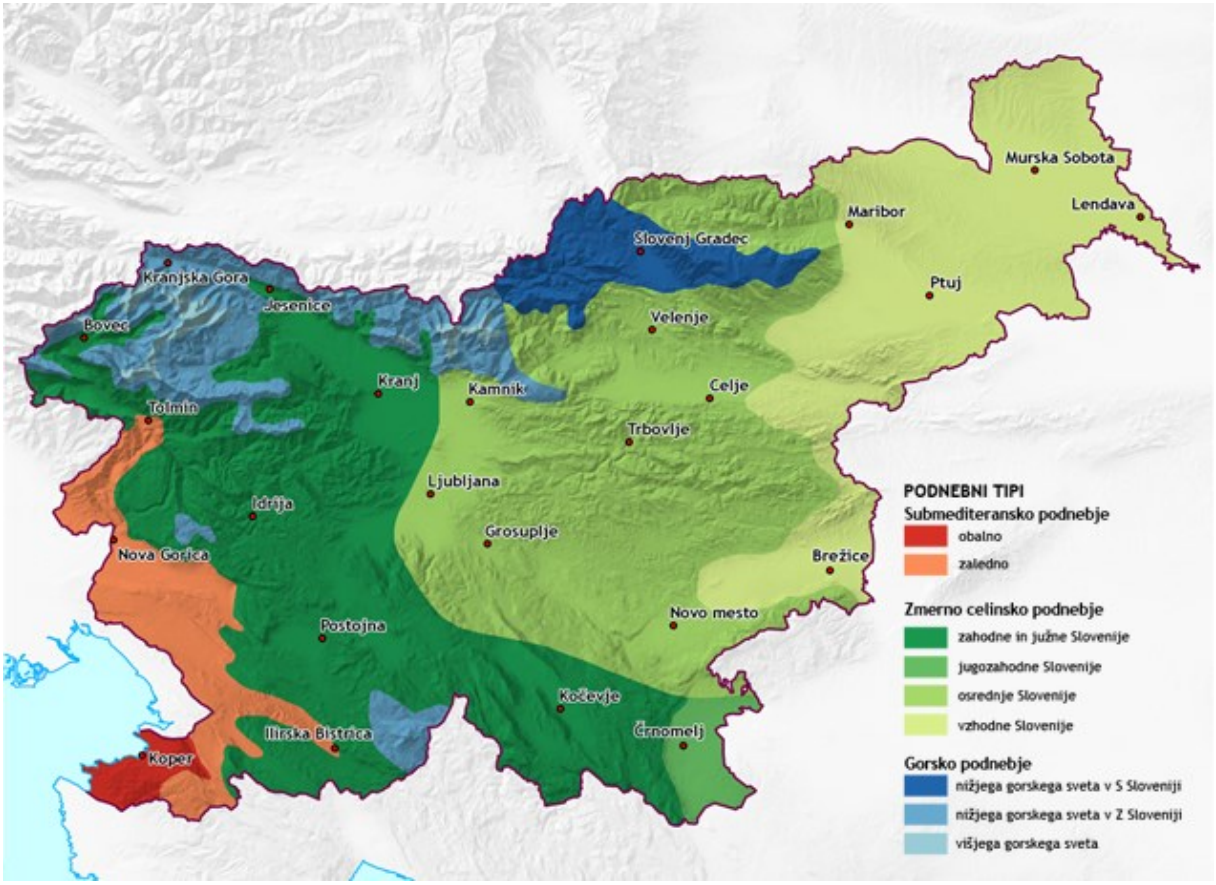
	NEZGODE Stevilo		NEZGODE Izplacilo		BOLEZNI Stevilo		BOLEZNI Izplacilo	
REGIJA	RMAE	RRMSE	RMAE	RRMSE	RMAE	RRMSE	RMAE	RRMSE
SLOVENIJA	0,7288	**0,7410**	0,8974	**0,9189**	1,0233	**1,0206**	1,0698	**1,0311**
PREDALPSKO	0,8650	**0,8691**	0,9911	**0,9853**	1,0425	**1,0300**	1,1280	**1,0649**
ALPSKO	0,9306	**0,9279**	1,0364	**1,0019**	1,0161	**1,0265**	1,0778	**1,0433**
PREDDINARSKO	0,9717	**0,9704**	1,0376	**1,0089**	1,0334	**1,0470**	1,1062	**1,0309**
DINARSKO	1,0237	**1,0086**	1,1100	**1,0533**	1,0797	**1,0511**	1,1046	**1,0381**
SUBPANONSKO	0,9236	**0,9165**	1,0178	**0,9975**	1,0132	**1,0204**	1,0497	**1,0263**
SUBMEDITERANSKO	0,9687	**0,9507**	1,0848	**1,0179**	1,0038	**1,0357**	1,1299	**1,0500**

5.7.6. Podnebni tipi

Razdelitev Slovenije je možna tudi glede na tipe podnebja. Glede na podatke geodetskega inštituta Slovenije [27] poznamo v Sloveniji tri tipe podnebja, ki jih je možno nadalje še podrobneje opredeliti:

- submediteransko podnebje:
  - obalno,
  - zaledno.
- Zmerno celinsko podnebje
  - zahodne in južne Slovenije,
  - jugovzhodne Slovenije,
  - osrednje Slovenije,
  - vzhodne Slovenije.
- Gorsko podnebje:
  - nižjega gorskega sveta v severni Sloveniji,
  - nižjega gorskega sveta v zahodni Sloveniji,
  - višjega gorskega sveta.

Razmejitev med podnebjema nižjega gorskega sveta zahodne Slovenije in visokega gorskega sveta je na zemljevidu (glej sliko 27) težje določljiva, zato sta bili za potrebe analize ti dve podnebji obravnavani kot enotno podnebje. Razdelitev je bila izvedena na podlagi tega, v kateri podnebni tip spada večina občine opazovanega kraja.



Slika 27: Razdelitev Slovenije glede na podnebne tipe (Vir: Geodetski inštitut Slovenije [27]).

Rezultati ocene so podani v tabeli 17. Iz tabele je razvidno, da so se tudi pri razrezu glede na podnebne tipe Slovenije ocene uspešnosti algoritma precej poslabšale. Tudi ta poskus razreza se ni izkazal za ugodnega.

Tabela 17: Uspešnost algoritma RandomForest pri razrezu na tipe podnebja.

	NEZGODE Stevilo		NEZGODE Izplacilo		BOLEZNI Stevilo		BOLEZNI Izplacilo	
REGIJA	RMAE	RRMSE	RMAE	RRMSE	RMAE	RRMSE	RMAE	RRMSE
SLOVENIJA	0,7288	**0,7410**	0,8974	**0,9189**	1,0233	**1,0206**	1,0698	**1,0311**
SUBMEDITERANSKO_OBALNO	0,9918	**1,0055**	1,1346	**1,0457**	1,0433	**1,0421**	1,1830	**1,0432**
SUBMEDITERANSKO_ZALEDNO	0,9758	**0,9778**	1,0909	**1,0172**	1,0292	**1,0394**	1,1007	**1,0653**
ZMERNO_ZAHOD_JUG_SLOVENIJE	0,9159	**0,9127**	1,0562	**1,0176**	1,0481	**1,0453**	1,0879	**1,0337**
ZMERNO_JUGOVZHOD_SLOVENIJE	1,0018	**1,0296**	1,1225	**1,0317**	1,1082	**1,0457**	1,0845	**1,0376**
ZMERNO_OSREDNJA_SLOVENIJA	0,8553	**0,8598**	0,9593	**0,9673**	1,0310	**1,0299**	1,0938	**1,0392**
ZMERNO_VZHOD_SLOVENIJE	0,9158	**0,9104**	0,9979	**0,9845**	1,0102	**1,0221**	1,0457	**1,0236**
GORSKO_NIZJE	1,0230	**1,0205**	1,1150	**1,0400**	1,0799	**1,0422**	1,0880	**1,0397**
GORSKO_VISJE	0,9944	**0,9874**	1,0960	**1,0376**	1,0340	**1,0382**	1,0771	**1,0565**

5.7.7. Podnebni tipi – osnovni

Kot zadnjega od poskusov lokalizacije smo naredili razrez zgolj na osnovne podnebne tipe, predstavljene v prejšnjem razdelku. Pri tem razbitju Slovenijo razdelimo na tri dele:

- submediteransko podnebje,
- zmerno celinsko podnebje,
- gorsko podnebje.

Pri tem načinu razdelitve večinski del Slovenije pride v zmerno celinsko podnebje. To varianto smo preizkusili prav zaradi obsega te razdelitve. Rezultati uspešnosti algoritma pri takšni razdelitvi Slovenije so prikazani v tabeli 18. Tudi pri tej razdelitvi je opaziti poslabšanje rezultatov. Kljub velikosti ene izmed regij je tudi pri tej ogromni regiji prišlo do poslabšanja rezultatov v primerjavi z rezultatom za ozemlje celotne Slovenije. To nakazuje, da operiramo s premajhnim številom dogodkov.

Tabela 18: Uspešnost algoritma RandomForest pri razrezu na osnovne tipe podnebja.

	NEZGODE Stevilo		NEZGODE Izplacilo		BOLEZNI Stevilo		BOLEZNI Izplacilo	
REGIJA	RMAE	RRMSE	RMAE	RRMSE	RMAE	RRMSE	RMAE	RRMSE
SLOVENIJA	0,7288	**0,7410**	0,8974	**0,9189**	1,0233	**1,0206**	1,0698	**1,0311**
SUBMEDITERANSKO_PODNEBJE	0,9799	**0,9614**	1,0805	**1,0190**	1,0105	**1,0360**	1,1285	**1,0447**
ZMERNO_PODNEBJE	0,7540	**0,7624**	0,9060	**0,9126**	1,0286	**1,0265**	1,0920	**1,0486**
GORSKO_PODNEBJE	0,9583	**0,9621**	1,0569	**1,0189**	1,0186	**1,0358**	1,0858	**1,0524**

6. Sklepne ugotovitve

Pri podajanju sklepnih ugotovitev pozornost najprej posvetimo pregledu postavljene metodologije. V nadaljevanju se skoncentriramo na pregled izpolnjenosti ciljev, zastavljenih v uvodnem delu.

6.1. Metodologija

Na podlagi postavljene metodologije je bil za potrebe magistrskega dela razvit spremljajoči program. Razviti program je napisan v programskem jeziku Java; za metode podatkovnega rudarjenja so uporabljene knjižnice orodja Weka. Z napisanim programom je bila podprta vsa v tem magistrskem delu opisana funkcionalnost. Program nam bo v prihodnosti omogočal predvsem lažje in hitrejše prilagajanje morebitnih sprememb in prilagoditev tukaj postavljene metodologije.

Žal se je zastavljena metodologija izkazala za bolj ali manj neuporabno pri uspešnosti napovedovanja škodnih dogodkov. Kakor smo videli, za napovedovanje bolezni metodologija ni ustrezna. To je bilo morebiti celo nekako pričakovano, saj smo opozorili, da je pri boleznih le v redkih primerih mogoče napovedati točen datum dogodka. Ne vemo vedno natančno, kdaj je bil začetek bolezni. Če ne vemo natančnega datuma dogodka, nam potem tudi vremenske razmere na tak okvirni datum ne povedo kaj dosti.

Tudi pri napovedovanju nezgod nam postavljena metodologija ni vrnila zadovoljivih rezultatov. V primeru napovedovanja števila nezgod je rezultat sicer še nekako sprejemljiv in s tem pogojno uporaben. V primeru napovedovanja povprečne višine izplačila za nezgode pa je napovedovanje neuporabno.

Izjalovil se je tudi poskus prostorske lokalizacije napovedi. Predpostavili smo, da bomo za manjša območja natančneje napovedovali dogodke. Vsi poskusi lokalizacije so se končali s slabšim rezultatom v primerjavi z rezultatom za celotno Slovenijo. Predpostavljamo da je vzrok premalo zajetih dogodkov.

Vzroke za neuspeh skušamo pojasniti v nadaljevanju.

6.2. Diagnostika vzrokov neuspeha

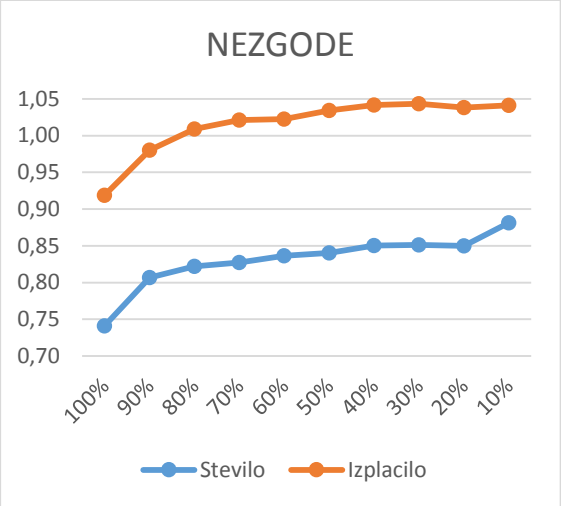
Že skozi postavljanje metodologije se nam nakazuje slutnja, da je v učno množico zajetih premalo dogodkov. To nekako nakazujejo tudi neuspešni poskusi lokalizacije, ko se nam pri manjšanju učne množice zmanjšuje tudi uspešnost napovedovanja.

Sklepamo, da učna množica zajema premalo dogodkov, da bi bil algoritem uspešen. To hipotezo preverimo. Z zmanjševanjem učne množice ob večkratnem poganjanju algoritma primerjamo rezultate uspešnosti napovedovanja za posamezne velikosti učnih množic. Če nam ob zmanjševanju učne množice pada tudi uspešnost napovedovanja, je to znak, da je učna množica premajhna.

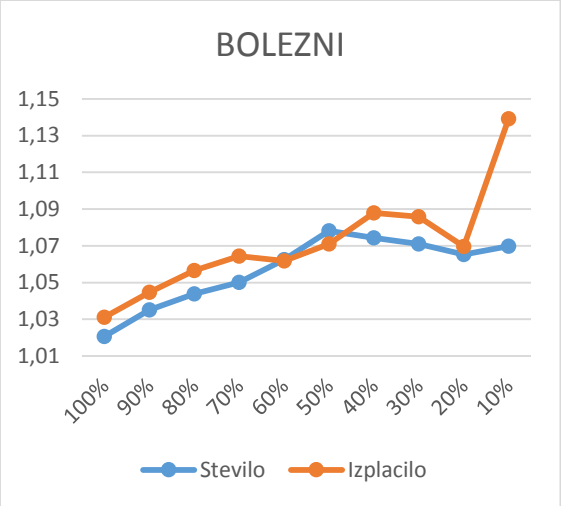
Za preverjanje zastavljene hipoteze smo učno množico korakoma zmanjševali za dodatnih 10 odstotkov originalne velikosti. Začeli smo z učno množico velikosti, kot je bila uporabljena pri ocenjevanju uspešnosti algoritma. V naslednjem koraku smo odstranili 10 odstotkov instanc iz množice in ponovili ocenjevanje. V naslednjem koraku smo zmanjšali začetno učno množico za 20 odstotkov. Korake smo ponavljali, dokler ni v učni množici ostalo samo še 10 odstotkov instanc originalne učne množice. Rezultati ocene uspešnosti takšnega zmanjševanja učne množice so predstavljeni v tabeli 19. Grafični prikaz padanja uspešnosti algoritma na podlagi RRMSE je za primer nezgod prikazan na sliki 28, za primer bolezni na sliki 29.

Tabela 19: RRMSE pri različnih velikostih učne množice.

RRMSE	NEZGODE		BOLEZNI	
Odstotek učne množice	Stevilo	Izplacilo	Stevilo	Izplacilo
100%	0,7410	0,9189	1,0206	1,0311
90%	0,8069	0,9801	1,0351	1,0447
80%	0,8223	1,0089	1,0439	1,0565
70%	0,8272	1,0211	1,0501	1,0645
60%	0,8363	1,0225	1,0625	1,0618
50%	0,8406	1,0344	1,0782	1,0710
40%	0,8503	1,0418	1,0744	1,0879
30%	0,8515	1,0433	1,0711	1,0859
20%	0,8501	1,0382	1,0652	1,0696
10%	0,8814	1,0415	1,0698	1,1392



Slika 28: RRMSE različnih velikosti množic za nezgode.



Slika 29: RRMSE različnih velikosti množic za bolezni.

Iz tabele 19 je razvidno, da je v primeru originalne velikosti učne množice uspešnost napovedovanja najboljša. Manjšanje učne množice povzroči slabšanje uspešnosti napovedovanja, saj se vrednost RRMSE z manjšanjem učne množice viša. Iz tega sklepamo, da bi algoritem v primeru večje učne množice pričakovano deloval bolje.

Kako doseči večjo učno množico, je vprašanje, ki se nam zastavlja ob tem spoznanju. V nadaljevanju predstavljamo nekaj variant, ki se porajajo kot možne rešitve:

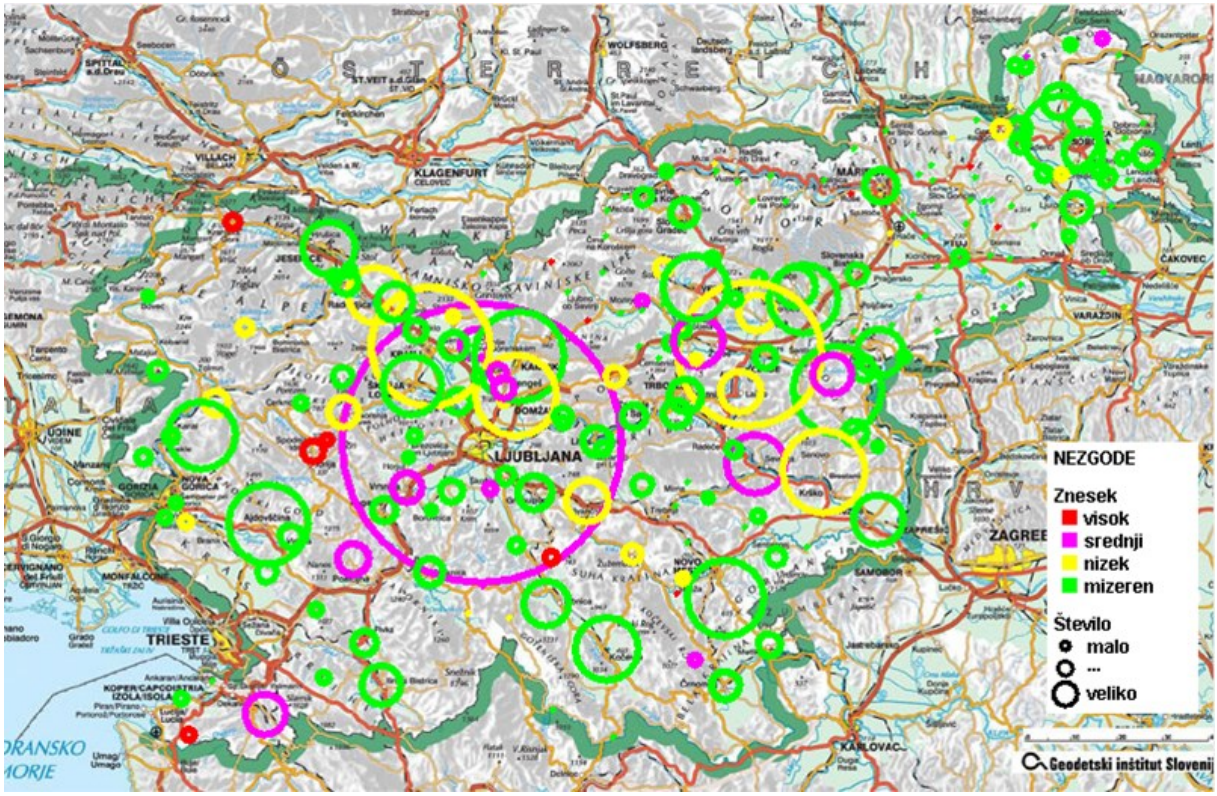
1. večjo učno množico dobimo, če povečamo časovno okno. To z obstoječimi podatki ne pride v poštev, saj smo ugotovili, da je uporabljena velikost okna s temi podatki optimalna. V primeru stabilizacije stanja v prihodnjih letih je povečanje časovnega okna dobra opcija.
2. Druga varianta je razširitev opazovanih dogodkov na dogodke sosednjih držav. Število odškodninskih dogodkov, ki so se zgodili v sosednjih državah, ni zanemarljivo. Pri tej rešitvi bi bilo nekaj težav s pridobivanjem vremenskih podatkov sosednjih držav, je pa to izvedljivo. Vprašanje je, če nam takšna razširitev model ne bi preveč zakomplicirala in s tem preprečila boljše rezultate.
3. Še ena izmed opcij za povečanje učne množice je razširitev zajema podatkov tudi na podatke drugih zavarovalnic v regiji. Tako bi precej povečali število opazovanih dogodkov. Ta opcija je malo verjetna, razen v primeru morebitnih združevanj zavarovalnic, kar pa je v zadnjem času pri nas dokaj pogost pojav.

6.3. Zemljevidi nezgod v Sloveniji skozi čas

Vse nezgode, ki se v Sloveniji pripetijo znotraj enega leta, želimo ponazoriti na karti. Za to bomo uporabili zemljevid Slovenije, saj bo tako lokacija predstavljena najbolj nazorno.

Če bi vsako nezgodo posebej zarisali na zemljevid, bi bili dogodki predstavljeni precej nenazorno, saj bi se točke v večini primerov prekrivale. Iz takšne karte ne bi bilo mogoče kaj dosti razbrati. Zaradi tega smo se odločili, da število nezgod ponazorimo z velikostjo kroga. Večji kot je krog, več nezgod se je na tistem mestu pripetilo. Ker je število krajev z nezgodami precejšnje, smo nivo prikazovanja dvignili na dogodke v občini. Prikazani krog ponazarja število nezgod v občini, katere glavni kraj obkroža krog. Tako je karta nezgod precej bolj berljiva.

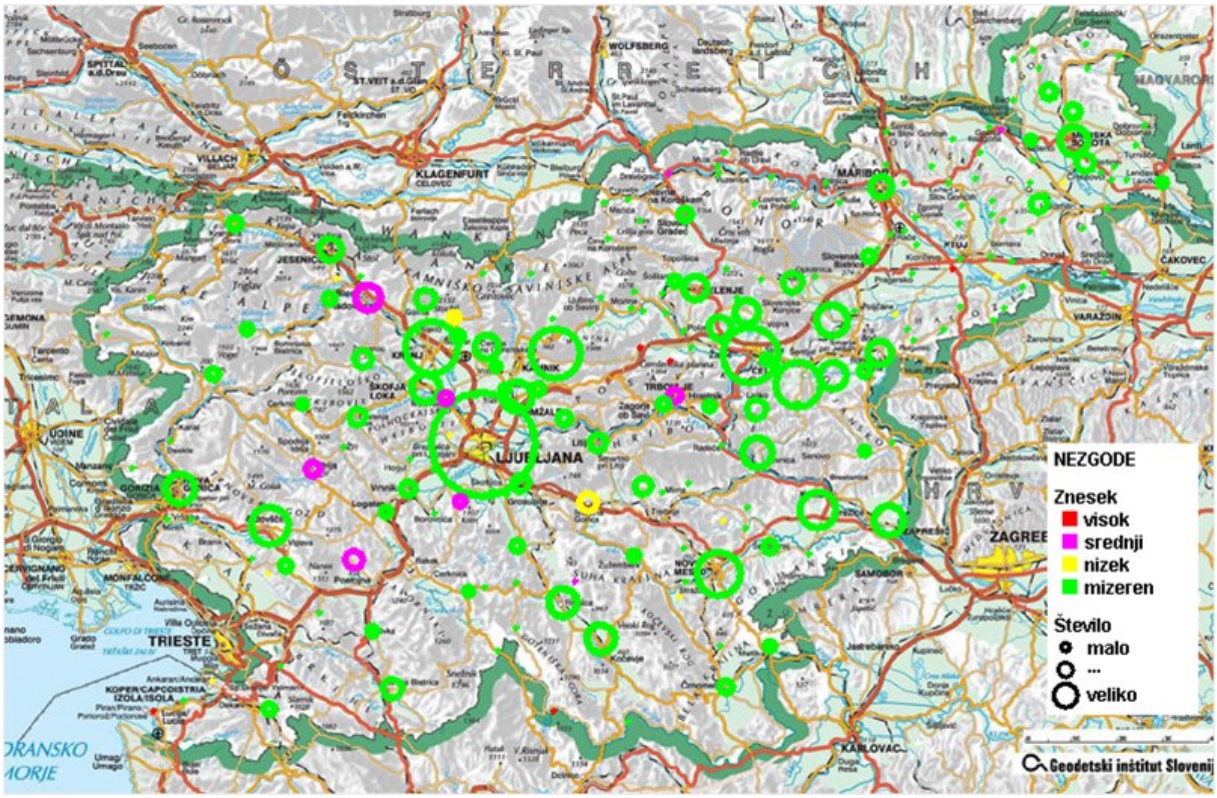
Poleg samega števila nezgod smo zemljevid uporabili tudi za prikaz povprečnih višin izplačanih nezgod v občini. Na istem zemljevidu smo to najlažje ponazorili z uporabo barv. Temnejše barve predstavljajo višji povprečni znesek izplačila, svetlejše barve ponazarjajo nižje zneske izplačila. Primer zemljevida s predstavljenimi nezgodami v letu 2007 vidimo prikazan na sliki 30.



Slika 30: Nezgode po slovenskih občinah v letu 2007.

Trendi nezgod

Spremljanje zemljevidov nezgod skozi vrsto let nam da dober vpogled v trende gibanja nezgod. Poleg zemljevida nezgod za leto 2007 (glej sliko30) si oglejmo še zemljevid nezgod za leto 2015 (glej sliko 31). Že na podlagi dveh let je mogoče napovedati trend nezgod za prihodnost. Izdelava zemljevidov je možna tudi za druge vzroke prijav škodnih dogodkov. Prilegajoča zemljevida za ugotavljanje trendov pri boleznih sta v prilogi C.



Slika 31: Nezgode po slovenskih občinah v letu 2015.

Iz prikazanih slik je videti, da se je velikost krogov v letu 2015 v primerjavi z letom 2007 na splošno bistveno zmanjšala. Predpostavljamo, da se ni zgodilo nič kaj dosti manj nezgod, ampak se to zmanjšanje bolj odraža v spremembi politike zavarovalnice. Nadzor ob prijavi nezgode se je v tem času bistveno poostril in je zato zdaj precej težje uveljavljati lažno nezgodo.

Poleg tega je v letu 2015 tudi barva krogov izrazito bolj svetla. Večinoma prevladuje zelena barva, kar pomeni, da se politika zaostrovanja nadzora kaže tudi v višini izplačanega nadomestila ob nezgodi. V letu 2015 so bili povprečni zneski izplačil bistveno nižji kot v preteklih obdobjih.

Kaže, da se je sprememba politike zavarovalnici obrestovala. Napredek v tehnologiji pripomore k učinkovitejšemu odkrivanju morebitnih prevar. V prihodnosti je tako pričakovati še bolj

učinkovito odkrivanje prevar, kar posledično ponovno vodi v nižje številke. Učinkovitejše odkrivanje prevar je dobrodošlo tudi med zavarovanci, saj to pomeni tudi nižje premije.

6.4. Najpomembnejši dejavniki za nastanek nezgode

Ko iščemo vzroke, ki privedejo do nezgode, moramo pomisliti tudi na vpliv vremenskih dejavnikov. Glede na opravljeno analizo [25] se samo na delovnem mestu pripeti 28 odstotkov nezgod zgolj zaradi neugodnih vremenskih razmer; vremenskega faktorja zato ne smemo zanemariti. Kateri vremenski parametri so tisti, ki so bolj odgovorni za nastanek nezgode, se zastavlja vprašanje.

Pri pregledu problema s številom nezgod vidimo, da se nezgode dogajajo sleherni dan. Že v razdelku 5.2 smo s pomočjo evalvacije ocenjevali vpliv atributov na določanje ciljnih atributov. Sklepamo, da, če ima nek parameter velik vpliv na določanje ciljnega atributa, ima tudi velik vpliv na sam pripetljaj nezgode. S pomočjo evalvacije smo v našem primeru na nek način videli, kolikšen vpliv ima na nastanek nezgode posamezen ocenjevani atribut.

Najpogostejši vzroki za nastanek nezgode so predstavljen z atributi, ki so se pri evalvaciji izkazali za najvplivnejše. Najbolj vplivni atributi so tisti, katerih ocena pri evalvaciji se pojavlja v samem vrhu tabele. V tabeli 20 povzemamo ocene atributov klasifikacijskega problema iz tabel 4 in 5. Prikazani so atributi, ki so bili ocenjeni vsaj s 33 % vrednosti najbolje ocenjenega atributa za evaluator.

Tabela 20: Najbolje ocenjeni atributi števila nezgod pri klasifikacijskem problemu.

razmerje informacijskega prispevka		**ReliefF**	
atribut	**ocena**	**atribut**	**ocena**
Rosa	0,06630	DanVTednu	0,03764
TrajanjeSoncnegaObsevanja	0,06008	LetniCas	0,01745
PovprecnaOblacnost	0,05983	LuninaMena	0,01704
Mesec	0,05053	Delovnik	0,01368
Teden	0,04753		
PovprecnaRelativnaVlaga	0,04634		
Dez	0,04598		
MaksimalnaTemperaturaZrakaNa2m	0,04146		
Rosenje	0,03924		
PovprecnaTemperaturaZrakaNa2m	0,03627		
PovprecnaHitrostVetra	0,03421		
SneznaOdeja	0,03222		
Meglica	0,03029		
DezSSnegom	0,02966		
SkupnaVisinaSnezneOdeje	0,02758		
Padavine	0,02439		

V tabeli 21 povzemamo ocene atributov regresijskega problema iz tabel 6 in 7. Tudi v tej tabeli so prikazani zgolj atributi, ki so bili ocenjeni vsaj s 33 % vrednosti najbolje ocenjenega atributa za evaluator.

Tabela 21: Najbolje ocenjeni atributi števila nezgod pri regresijskem problemu.

pričakovana razlika variance		**RReliefF**	
atribut	**ocena**	**atribut**	**ocena**
DanVTednu	0,00069	Sneg	0,00428
LetniCas	0,00060	MinimalnaTemperaturaZrakaNa5cm	0,00307
LuninaMena	0,00059	SkupnaVisinaSnezneOdeje	0,00284
Rosa	0,00042	Grmenje	0,00266
Dez	0,00041	Delovnik	0,00247
Padavine	0,00040	DnevnaKolicinaPadavin	0,00230
SneznaOdeja	0,00040	MaksimalnaTemperaturaZrakaNa2m	0,00225
PlohaDezja	0,00040	VisinaNovozapadlegaSnega	0,00223
DezKiZmrzuje	0,00040	Nevihta	0,00213
Sneg	0,00040	Dez	0,00212
Slana	0,00040	LuninaMena	0,00207
Rosenje	0,00040	PovprecnaTemperaturaZrakaNa2m	0,00194
Delovnik	0,00040	LetniCas	0,00172
Meglica	0,00040	MinimalnaTemperaturaZrakaNa2m	0,00157
Grmenje	0,00039	DezKiZmrzuje	0,00144
Ivje	0,00039		
Nevihta	0,00039		
ViharniVeter	0,00039		
TrdoIvje	0,00039		

V tabelah 20 in 21 smo z zeleno barvo obarvali atribute, ki se pojavijo vsaj pri treh od štirih evalvatorjev. Pričakovana razlika variance ocenjuje samo diskretne atribute, kar je eden izmed razlogov, zakaj zadostuje že pojav pri treh evalvatorjih. Atributi, ocenjeni z manj kot 33 % ocene najvišje ocenjenega atributa, so bili pri tem razmisleku izpuščeni, saj je njihov vpliv na ciljni atribut premajhen. Atribute smo obtežili glede na njihovo oceno in kot najpomembnejše dejavnike za nastanek nezgode identificirali naslednje atribute:

- LetniCas,
- Dez,
- LuninaMena,
- Delovnik.

Enako smo naredili za problem nastanka bolezni. Pri boleznih smo kot najpomembnejši dejavnik za nastanek bolezni identificirali atribut LetniCas. Rezultati za bolezni so predstavljeni v dodatku D.

6.4.1. Dejavniki ekstremnih nezgod

Izraz ekstremna nezgoda je morda nekoliko pretiran, vendar smo ga mi uporabili zavoljo nazornosti odstopanja vrednosti ciljnega atributa od večine preostalih primerov. O ekstremni nezgodi govorimo v primeru nezgode, katere rezultat so hujše posledice. Takšne nezgode po navadi zaznamo na podlagi visokega izplačila odškodnine. Kot drug primer takšnih ekstremnih nezgod obravnavamo primer, ko se v enem dnevu pripeti ogromno število nezgod - tolikšno število nezgod, da bistveno odstopajo od siceršnjega dnevnega povprečja.

Predpostavimo, da ta odstopanja predstavljajo zgornje tri odstotke vrednosti pri opazovanih dneh. Na podlagi te predpostavke ovrednotimo atribute po enaki metodi, kot smo jo uporabili v prejšnjem razdelku. Zgolj nad množico teh treh odstotkov podatkov ponovimo evalvacijo atributov za problem števila nezgod. Zaradi problema regresijske narave uporabimo evalvacijo s pričakovano razliko varianc in algoritem RReliefF. Rezultati vrednotenja atributov so prikazani v tabeli 22.

Tabela 22: Najbolje ocenjeni atributi števila ekstremnih nezgod pri regresijskem problemu.

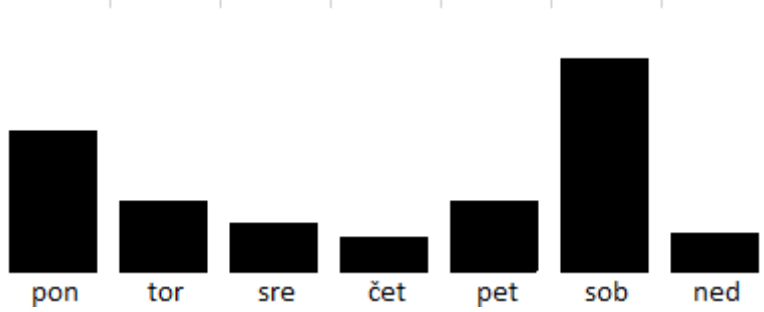
pričakovana razlika variance		**RReliefF**	
atribut	**ocena**	**atribut**	**ocena**
DanVTednu	0,00021	DanVTednu	0,05048
LetniCas	0,00018	LetniCas	0,04403
LuninaMena	0,00018	MinimalnaTemperaturaZrakaNa5cm	0,02657
Delovnik	0,00012	Mesec	0,02068
		Teden	0,01454
		MinimalnaTemperaturaZrakaNa2m	0,01381
		PovprecnaTemperaturaZrakaNa2m	0,01243
		PovprecnaOblacnost	0,01209
		LuninaMena	0,00817
		PovprecnaHitrostVetra	0,00487
		Delovnik	0,00283
		MaksimalnaTemperaturaZrakaNa2m	0,00020

V tabeli 22 smo z rdečo barvo označili vrednosti, pri katerih je ocena atributa nižja od 33 % najbolje ocenjenega atributa. Ti atributi niso prišli v poštev pri določanju najpomembnejših atributov. Kot najpomembnejše atribute za nastanek ekstremnih nezgod smo identificirali naslednje atribute:

- DanVTednu,
- LetniCas.

Kot najbolj markanten se izkaže atribut, ki predstavlja dan v tednu. Na sliki 32 vidimo, da se največje število ekstremnih nezgod pripeti v soboto. Za to je kriv vpliv vikenda in tudi porast

zanimanja za športne aktivnosti. Ljudje imajo ob sobotah večinoma prosto in to izkoristijo za razne aktivnosti. Zadnji dan v vikendu je za razliko od sobote bistveno manj nezgod, saj naj bi bile nedelje namenjene počitku. Drugi najvplivnejši parameter to tezo le še potrjuje. Večino nezgod zaznamo v zimskem času, ko vemo, da se dosti ljudi ukvarja z zimskimi športi.



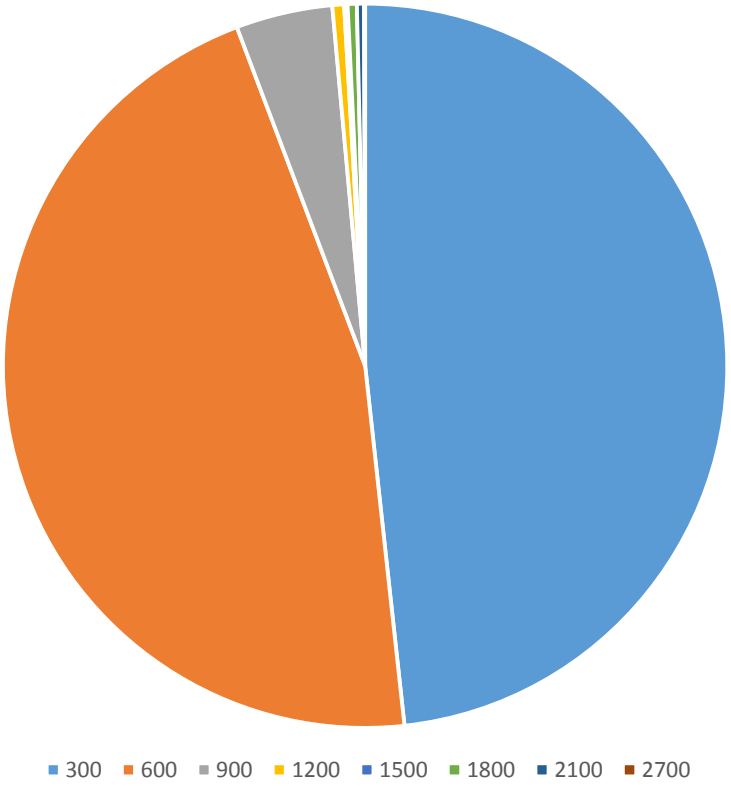
Slika 32: Razporeditev števila ekstremnih nezgod v tednu.

Podobno zgodbo imamo pri boleznih. Tabelo vrednotenja atributov pri ekstremnih boleznih podajamo v prilogi E. Tudi pri ekstremnih boleznih dobimo kot najvplivnejše za nastanek bolezni enake atribute.

6.5. Vpliv nadmorske višine na nezgodne dogodke

Za analizo podatkov o vplivu nadmorske višine smo uporabili podatke od začetka leta 2010 do konca leta 2015. Pri starejših nezgodah zaradi drugačnega zajema podatkov ni mogoče dovolj natančno določiti lokacije nezgode, s tem pa tudi ne prave nadmorske višine za lokacijo.

Statistični pregled podatkov nam pokaže, da število nezgod z naraščanjem nadmorske višine pada. To je po pričakovanjih, saj se večina ljudi zadržuje v nižinah in se zato tam zgodi največ nezgod. Pregled števila nezgod glede na nadmorsko višino podajamo na sliki 33 in v pripadajoči tabeli 23.



Slika 33: Porazdelitev števila nezgod glede na nadmorsko višino.

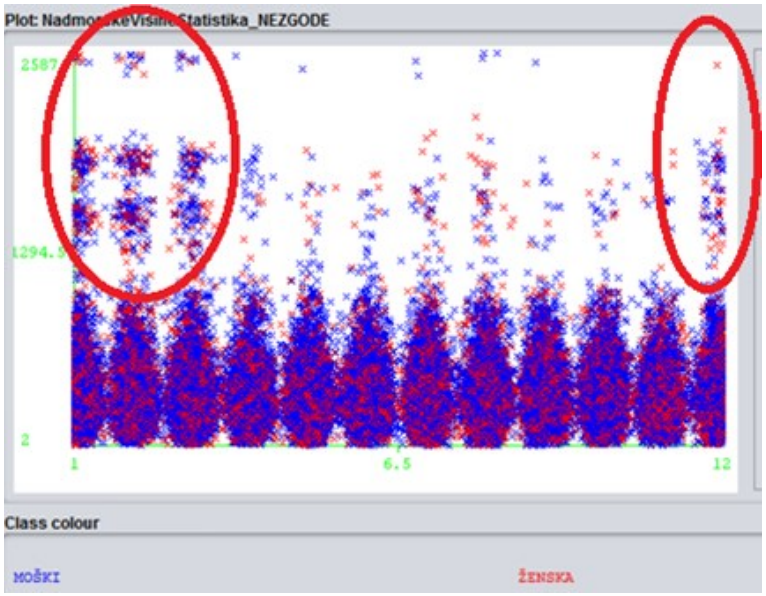
Tabela 23: Porazdelitev števila nezgod glede na nadmorsko višino.

Nadmorska višina (m)	Odstotek nezgod (%)
300	48,27
600	45,97
900	4,31
1200	0,52
1500	0,15
1800	0,41
2100	0,32
2700	0,05

Pri pregledu nezgod zaznamo pričakovane vzorce pojavljanja, ki jih je mogoče preprosto razložiti. Vmes zasledimo tudi kakšno zanimivost, ki bi morda bila vredna nadaljnje raziskave. V nadaljevanju predstavljamo nekaj izmed teh odkritih vzorcev, ki smo jih odkrili pri analizi podatkov.

6.5.1. Zima pomeni več nezgod v hribih

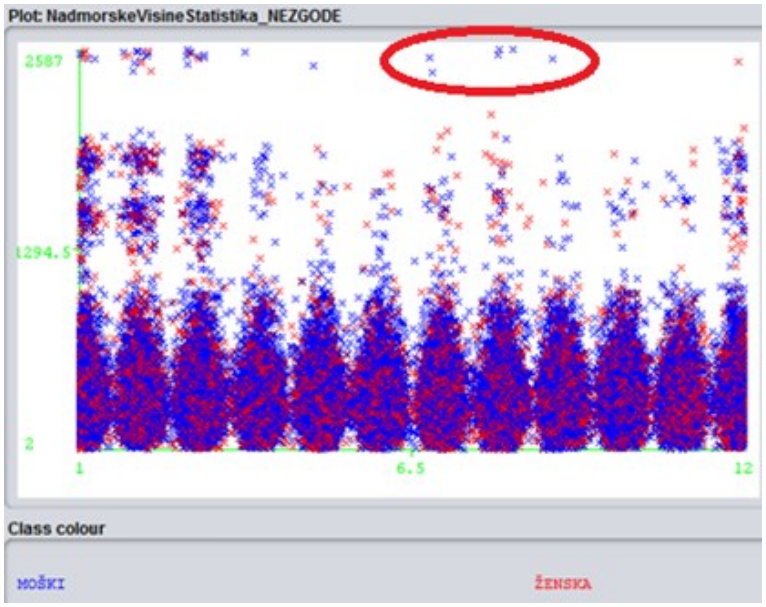
Iz podatkov je zaznati, da se v zimskih mesecih v hribih zgodi bistveno več nezgod. Pojav je pričakovan, saj so pozimi v hribih aktualni zimski športi, pri katerih je verjetnost za nezgodo precej povišana. Na sliki 34 vidimo prikazane nezgode po mesecih glede na spol. Iz slike je razbrati, da se pri višini nad 1200 metrov v zimskih mesecih zgodi več nezgod. Spol pri tem ne igra vidne vloge.



Slika 34: Povečanje števila nezgod v hribih zaradi zimskih športov.

6.5.2. Ali v visokogorje zahajajo predvsem moški?

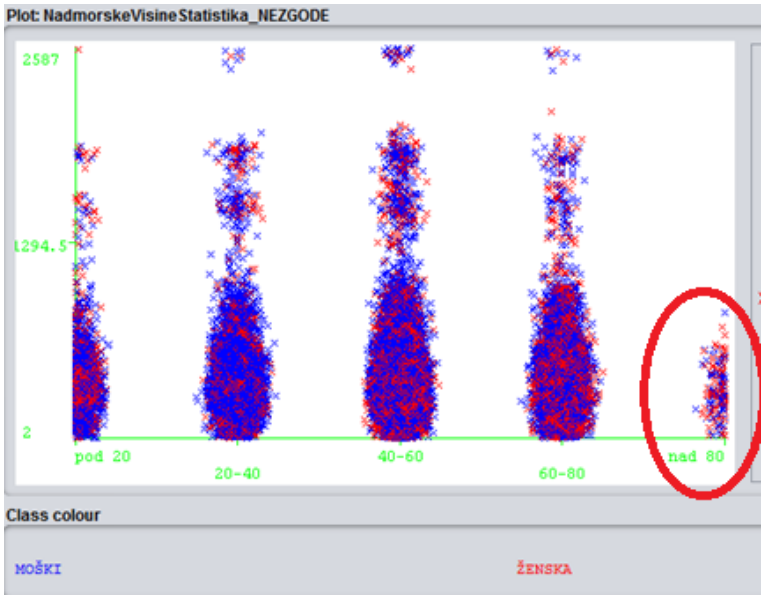
Zanimivo je, da v poletnih mesecih med nezgodami v visokogorju ne zasledimo nezgod med ženskami. Na sliki 35 je videti, da so pri nadmorski višini nad 2300 metov v poletnih mesecih zgolj nezgode med moškimi. Pojav morda razložimo s trditvijo, da se za obisk visokogorja odloča manj žensk. Možno je tudi, da je za visokogorje vzorec podatkov premajhen za postavljanje takšnih tez. Ena izmed teorij zagovarja, da imajo ženske nižje težišče in bi naj zato bile bolj stabilne. Morda je nekaj tudi na tem, da naj bi bile ženske na splošno bolj previdne.



Slika 35: Nezgode moških v visokogorju v poletnih mesecih.

6.5.3. Starostniki neradi zapuščajo dom

Znano je dejstvo, da so starostniki manj aktivni. Če gledamo starostno skupino nad 80 let, sicer zajamemo manjšo skupino zavarovancev, vendar menimo, da je dovolj zgovorno dejstvo, da se starostnikom nobena nezgoda ni pripetila na nadmorski višini nad 1000 metrov (glej sliko 36). To si razlagamo s tem, da se starostniki verjetno raje zadržujejo v bližini svojega doma. Ker niso toliko aktivni, ne zahajajo pogosto v hribe.

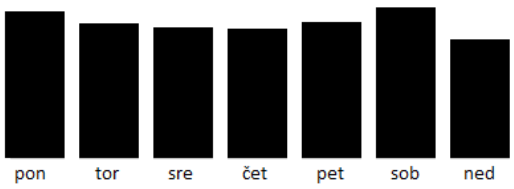


Slika 36: Nezgode starostnikov glede na nadmorsko višino.

6.5.4. Ob vikendih je več nezgod v hribih

Opazen vpliv na nezgode v hribih ima vikend. Ob vikendih so ljudje večinoma prosti in se zato bolj lotevajo raznih aktivnosti v hribih. V zimskem času to obsega zimske športe, v poletnih mesecih hojo v hribe. Populacija v hribih se za vikende poveča in je zato večja verjetnost nezgod. Na sliki 37 je videti, da je nad 1000 metri za vikend porast prijavljenih nezgod. Za primerjavo dodajamo še normalno razporeditev nezgod za vse nadmorske višine (glej sliko 38).

Slika 37: Nezgode nad 1000 m glede na dneve v tednu.



Slika 38: Vse nezgode glede na dneve v tednu.

7. Zaključki

V magistrskem delu je predstavljen razvoj metodologije, ki naj bi omogočala napovedovanje odškodninskih zahtevkov. To napovedovanje se izvaja na podlagi preteklih prijav škodnih dogodkov v soodvisnosti od vremenskih podatkov. Z uporabo knjižnic orodja Weka smo na podlagi tehnik prostorsko-časovnega rudarjenja razvili javanski program za podporo temu napovedovanju.

Žal se je izkazalo, da zastavljena metodologija ni primerna za napovedovanje naših primerov. Ugotovljeno je bilo, da bi boljše rezultate dobili, če bi bilo v obravnavo zajetih več škodnih dogodkov. Predstavljenih je bilo nekaj načinov za dosego tega cilja. Poleg tega smo ugotovili, da se metodologija nekoliko bolje obnese pri napovedovanju nezgod kakor pri napovedovanju bolezni. To nam da slutiti, da bi morali različne vzroke prijave obravnavati ločeno. Vsak vzrok prijav ima svoje posebnosti in bi ga bilo smiselno samostojno obravnavati.

Spremembe v poslovanju zavarovalnice kažejo, da je trenutno nekoliko neugoden čas za postavljanje takšne metodologije. Opaziti je bilo, da se je med letoma 2000 in 2015 trend odškodninskih zahtevkov opazovanih vzrokov obrnil v popolnoma drugo smer. Čeprav smo obravnavali le leta od 2010 do 2015, je bilo obračanje trenda še vedno opazno. Poskusimo razložiti obračanje trenda, pri katerem gre za zniževanje števila prijav in višine povprečnega izplačila odškodnine. Vse večji in strožji nadzor nad prijavami je razlog, da do tega prihaja. V zadnjih letih je bilo v to vloženo precej truda, zato lahko v prihodnosti pričakujemo umiritev tega trenda. Dogajanje je potrebno spremljati z mislijo, da je to obračanje trenda v interesu vseh. Zavarovalnice si na ta račun nižajo obratovalne stroške, zavarovanci pa v končni fazi lahko računamo na ugodnejše zavarovalne premije.

V magistrskem delu je bila ugotovljena tudi pomanjkljivost glede opisnega vnosa kraja dogodka. Podali smo predlog za predelavo vnosne maske na način, da se bo kraj izbiral iz šifranta krajev v Sloveniji. Za primer dogodkov v tujini naj ostane tudi opisno polje. Za nadaljnje analize je ta prilagoditev nujno potrebna.

Osnovni cilj magistrskega dela je dosežen, saj je bila metodologija izdelana. Namen lokalizacije dogodkov sicer ni bil izpolnjen, a smo pokazali, da nam je to preprečilo majhno število opazovanih dogodkov. V magistrskem delu so bili izpolnjeni tudi vsi drugi v uvodu zastavljeni cilji. Izdelani so bili zemljevidi za prikaz odškodninskih zahtevkov v teku let. Analizirani so

bili dejavniki, ki vplivajo na nastanek nezgod. Pripravili smo analizo vpliva nadmorske višine na odškodninske zahtevke nezgod.

Ugotavljamo, da kljub navideznemu neuspehu, trud ni bil zaman. Za napovedovanje nezgod je metodologija kljub vsemu delno uporabna, jo je pa mogoče še dodelati in jo izpopolniti. Mogoče je tudi, da se uporabnost postavljene metodologije pokaže šele v prihodnosti.

7.1. Nadaljnje delo

Glede na to, da v tem magistrskem delu postavljena metodologija še ni povsem pripravljena za uporabo, se bo delo na njej še nadaljevalo. Še vedno nismo izčrpali vseh možnosti.

Kakor smo že omenili, bi bil prvi smiselni nadaljnji korak ločitev posameznih vzrokov odškodninskih zahtevkov. Smiselno bi se bilo najprej omejiti zgolj na nezgode in ostale vzroke prihraniti za kasnejše analize. V prvi fazi bi se osredotočili le na bolj obetavne nezgode, bolezni pa bi trenutno pustili ob strani.

Še vedno ostaja vrsta nepreizkušenih metod, ki bi se jih prav tako splačalo preizkusiti. Poleg že uporabljenih bolj osnovnih metod bi veljalo preizkusiti tudi učinkovitost nekaterih zahtevnejših metod. Tu imamo v mislih metodo podpornih vektorjev (angl. SVM – support vector machine), mogoče pa poskusimo tudi z nevronskimi mrežami.

V točki 6.2 smo se že dotaknili nekaterih variant, s katerimi bi bilo možno povečati število dogodkov učne množice. Na tem mestu bi morda omenil samo še nekaj. Namesto da enoto časovne skale razširjamo, bi šli v popolnoma drugo stran in bi dan raje poskusili skrčiti. Dan bi razdelili na več manjših enot. Kot rešitev se nam sama nakazuje razdelitev dneva na osem ur, saj na nekaterih vremenskih postajah merjenje določenih vremenskih spremenljivk poteka trikrat dnevno.

Priloge

A. Ocena atributov za bolezni

A.1. Razmerje informacijskega prispevka

Tabela 24: Ocena atributov z razmerjem informacijskega prispevka pri boleznih.

Število bolezni		**Povprečno izplačilo bolezni**	
atribut	**ocena**	**atribut**	**ocena**
Teden	0,01937	Teden	0,01401
Mesec	0,01717	LetniCas	0,00507
SneznaOdeja	0,01660	DanVTednu	0,00308
MinimalnaTemperaturaZrakaNa2m	0,01642	Delovnik	0,00164
MinimalnaTemperaturaZrakaNa5cm	0,01543	LuninaMena	0,00093
MaksimalnaTemperaturaZrakaNa2m	0,01538		
PovprecnaTemperaturaZrakaNa2m	0,01410		
LetniCas	0,00730		
DanVTednu	0,00350		
LuninaMena	0,00242		
Delovnik	0,00022		

A.2. ReliefF

Tabela 25: Ocena atributov z ReliefF pri boleznih.

Število bolezni		**Povprečno izplačilo bolezni**	
atribut	**ocena**	**atribut**	**ocena**
Mesec	0,00494	Teden	0,00147
Delovnik	0,00328	LuninaMena	0,00123
LetniCas	0,00287	MinimalnaTemperaturaZrakaNa5cm	0,00116
Teden	0,00286	PovprecnaHitrostVetra	0,00102
PovprecnaTemperaturaZrakaNa2m	0,00240	Mesec	0,00078
VisinaNovozapadlegaSnega	0,00143	Delovnik	0,00067
DanVTednu	0,00140	MaksimalnaTemperaturaZrakaNa2m	0,00043
PovprecnaRelativnaVlaga	0,00119	DanVTednu	0,00032
MinimalnaTemperaturaZrakaNa2m	0,00091	LetniCas	0,00024
PovprecenZracniTlak	0,00034	PovprecnaTemperaturaZrakaNa2m	0,00016
PovprecnaHitrostVetra	0,00029	MinimalnaTemperaturaZrakaNa2m	0,00004
MinimalnaTemperaturaZrakaNa5cm	0,00026		
DnevnaKolicinaPadavin	0,00017		
PovprecnaOblacnost	0,00015		
LuninaMena	0,00012		
MaksimalnaTemperaturaZrakaNa2m	0,00005		
SkupnaVisinaSnezneOdeje	0,00004		

A.3. Pričakovana razlika variance

Tabela 26: Ocena atributov s pričakovano razliko variance pri boleznih.

Število bolezni		**Povprečno izplačilo bolezni**	
atribut	**ocena**	**atribut**	**ocena**
DanVTednu	0,00003	DanVTednu	0,00171
LetniCas	0,00003	LetniCas	0,00150
LuninaMena	0,00003	LuninaMena	0,00150
SneznaOdeja	0,00002	DezKiZmrzuje	0,00100
PlohaDezja	0,00002	Rosa	0,00100
Rosa	0,00002	Dez	0,00100
Rosenje	0,00002	PlohaDezja	0,00100
Grmenje	0,00002	Padavine	0,00100
Sneg	0,00002	Sneg	0,00100
Nevihta	0,00002	Delovnik	0,00100
Slana	0,00002	Slana	0,00100
Dez	0,00002	Rosenje	0,00100
Meglica	0,00002	Nevihta	0,00100
DezKiZmrzuje	0,00002	Grmenje	0,00100
Ivje	0,00002	Meglica	0,00100
Delovnik	0,00002	SneznaOdeja	0,00100
ViharniVeter	0,00002	Ivje	0,00100
Padavine	0,00002	TrdoIvje	0,00100
TrdoIvje	0,00002		

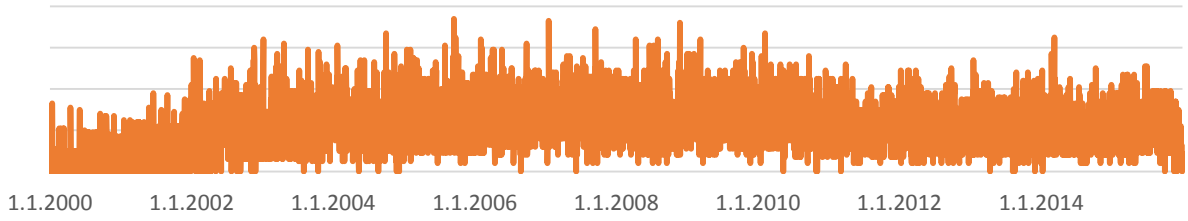
A.4. RReliefF

Tabela 27: Ocena atributov z RReliefF pri boleznih.

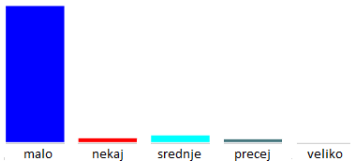
Število bolezni		**Povprečno izplačilo bolezni**	
atribut	**ocena**	**atribut**	**ocena**
ViharniVeter	0,00756	LuninaMena	0,01500
Rosenje	0,00585	PovprecnaOblacnost	0,00431
MocanVeter	0,00280	PovprecenZracniTlak	0,00308
MaksimalnaTemperaturaZrakaNa2m	0,00243	PovprecnaRelativnaVlaga	0,00273
Mesec	0,00205	PovprecnaHitrostVetra	0,00273
MinimalnaTemperaturaZrakaNa2m	0,00199	Delovnik	0,00248
Delovnik	0,00173	DnevnaKolicinaPadavin	0,00189
SkupnaVisinaSnezneOdeje	0,00168	LetniCas	0,00169
Dez	0,00166	Mesec	0,00159
Teden	0,00166	PovprecnaTemperaturaZrakaNa2m	0,00155
DanVTednu	0,00165	DanVTednu	0,00138
MinimalnaTemperaturaZrakaNa5cm	0,00165	MaksimalnaTemperaturaZrakaNa2m	0,00103
PovprecnaHitrostVetra	0,00156	MinimalnaTemperaturaZrakaNa2m	0,00060
PovprecnaTemperaturaZrakaNa2m	0,00131	MinimalnaTemperaturaZrakaNa5cm	0,00052
PovprecnaOblacnost	0,00127	Teden	0,00044
DnevnaKolicinaPadavin	0,00109		
PovprecenZracniTlak	0,00096		
VisinaNovozapadlegaSnega	0,00050		
PovprecnaRelativnaVlaga	0,00048		
LetniCas	0,00039		
LuninaMena	0,00033		
TrajanjeSoncnegaObsevanja	0,00030		

B. Razporeditev razredov za bolezni

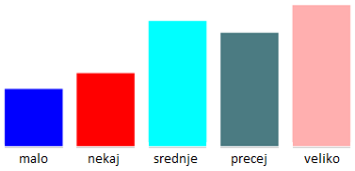
B.1. Število prijav bolezni od leta 2000 do leta 2015



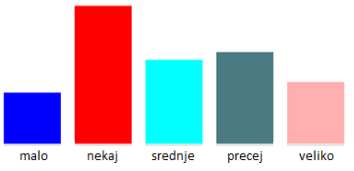
Slika 39: Število prijavljenih bolezni od leta 2000 do leta 2015.



Slika 40: Število bolezni v letu 2000.

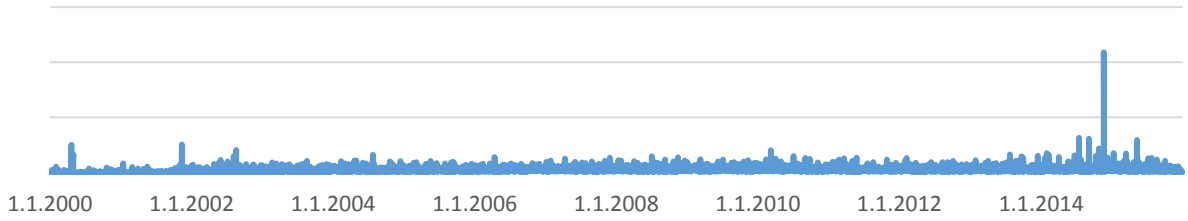


Slika 41: Število bolezni v letu 2008.



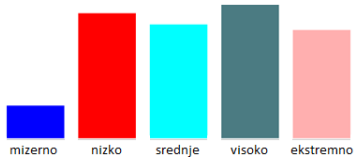
Slika 42: Število bolezni v letu 2015.

B.2. Povprečna višina izplačane bolezni od leta 2000 do leta 2015

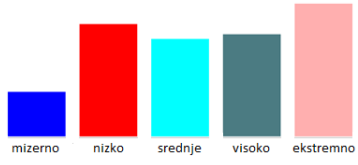


Slika 43: Povprečna višina izplačane bolezni od leta 2000 do leta 2015.

Slika 44: Povprečno izplačilo bolezni v letu 2000.

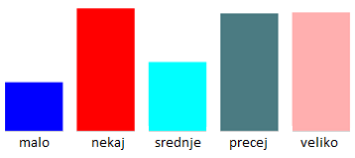


Slika 45: Povprečno izplačilo bolezni v letu 2008.

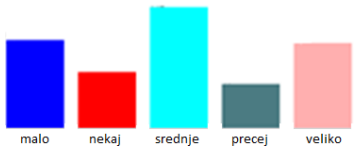


Slika 46: Povprečno izplačilo bolezni v letu 2015.

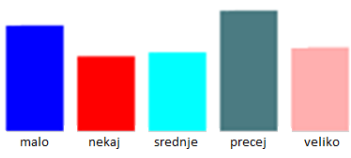
B.3. Število prijav bolezni od leta 2010 do leta 2015



Slika 47: Število bolezni v letu 2010.

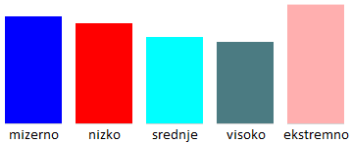


Slika 48: Število bolezni v letu 2013.

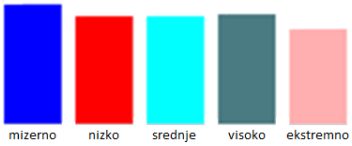


Slika 49: Število bolezni v letu 2015.

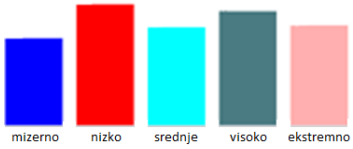
B.4. Povprečna višina izplačane bolezni od leta 2010 do leta 2015



Slika 50: Povprečno izplačilo bolezni v letu 2010.

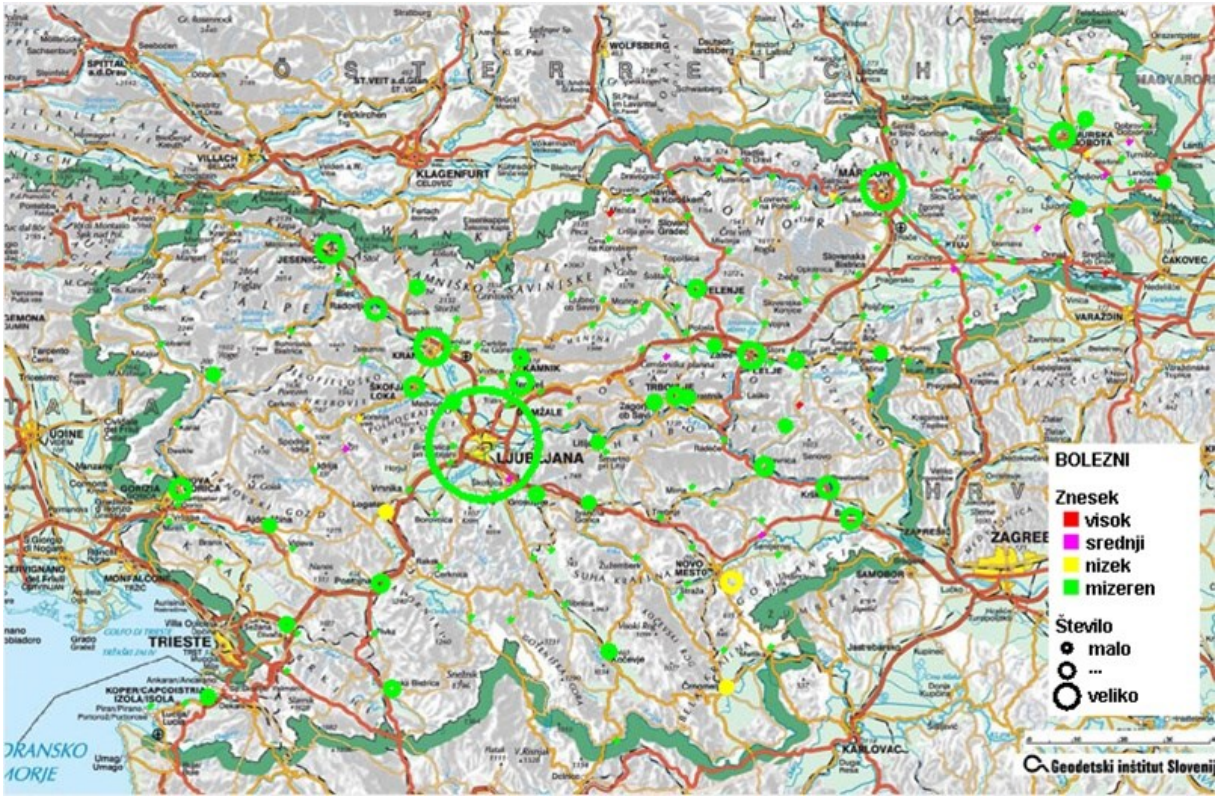


Slika 51: Povprečno izplačilo bolezni v letu 2013.

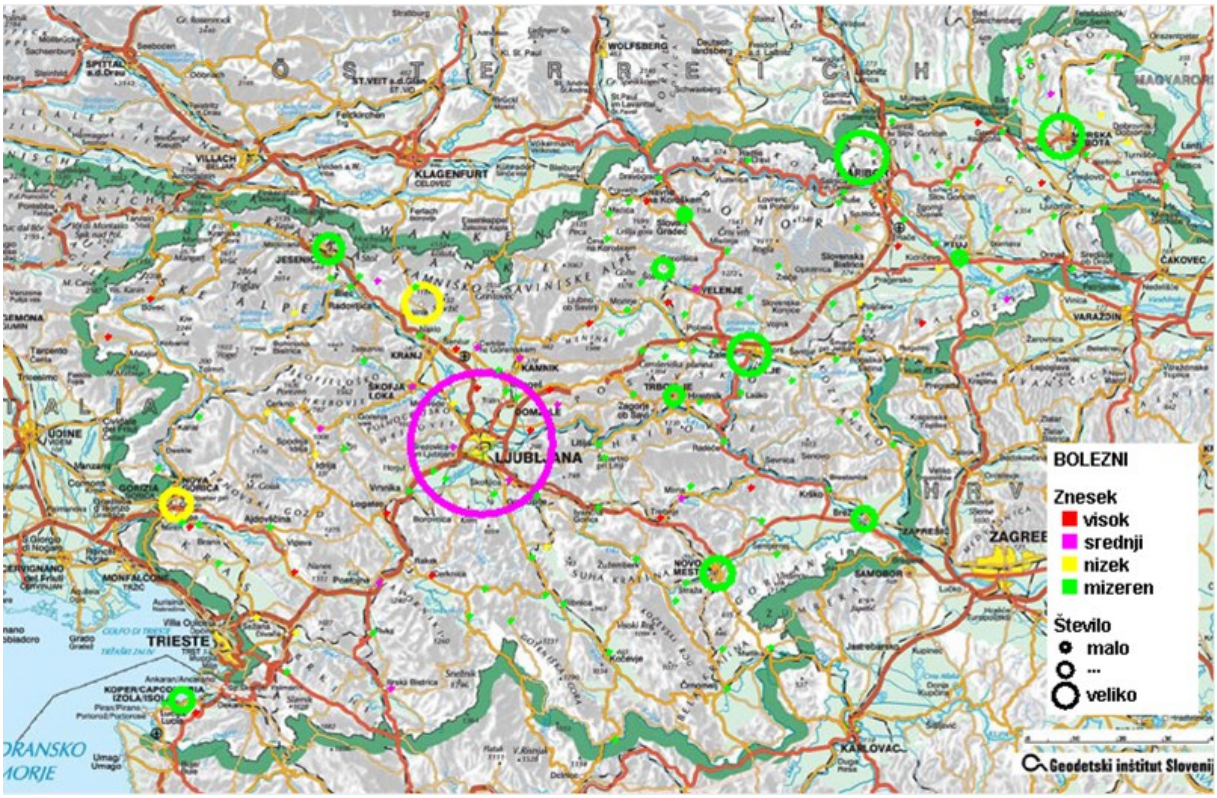


Slika 52: Povprečno izplačilo bolezni v letu 2015.

C. Trendi nastajanja bolezni



Slika 53: Bolezni po slovenskih občinah v letu 2007.



Slika 54: Bolezni po slovenskih občinah v letu 2015.

D. Najpomembnejši dejavniki za nastanek bolezni

Tabela 28: Najbolje ocenjeni atributi števila bolezni pri klasifikacijskem problemu.

razmerje informacijskega prispevka		ReliefF	
atribut	**ocena**	**atribut**	**ocena**
Teden	0,01937	Mesec	0,00494
Mesec	0,01717	Delovnik	0,00328
SneznaOdeja	0,01660	LetniCas	0,00287
MinimalnaTemperaturaZrakaNa2m	0,01642	Teden	0,00286
MinimalnaTemperaturaZrakaNa5cm	0,01543	PovprecnaTemperaturaZrakaNa2m	0,00240
MaksimalnaTemperaturaZrakaNa2m	0,01538		
PovprecnaTemperaturaZrakaNa2m	0,01410		
LetniCas	0,00730		

Tabela 29: Najbolje ocenjeni atributi števila nezgod pri regresijskem problemu.

pričakovana razlika variance		RReliefF	
atribut	**ocena**	**atribut**	**ocena**
DanVTednu	0,00003	ViharniVeter	0,00756
LetniCas	0,00003	Rosenje	0,00585
LuninaMena	0,00003	MocanVeter	0,00280
SneznaOdeja	0,00002		
PlohaDezja	0,00002		
Rosa	0,00002		
Rosenje	0,00002		
Grmenje	0,00002		
Sneg	0,00002		
Nevihta	0,00002		
Slana	0,00002		
Dez	0,00002		
Meglica	0,00002		
DezKiZmrzuje	0,00002		
Ivje	0,00002		
Delovnik	0,00002		
ViharniVeter	0,00002		
Padavine	0,00002		
TrdoIvje	0,00002		

E. Evaluacija ekstremnih bolezni

Tabela 30: Najbolje ocenjeni atributi števila ekstremnih bolezni pri regresijskem problemu.

pričakovana razlika variance		RReliefF	
atribut	**ocena**	**atribut**	**ocena**
DanVTednu	0,00001	DanVTednu	0,05722
LuninaMena	0,00001	LetniCas	0,03201
LetniCas	0,00001	LuninaMena	0,01591
Delovnik	0,00001	MinimalnaTemperaturaZrakaNa2m	0,00618
		PovprecnaTemperaturaZrakaNa2m	0,00556
		Mesec	0,00512
		MaksimalnaTemperaturaZrakaNa2m	0,00500
		Teden	0,00436
		Delovnik	0,00339

Literatura

[1] Miran Borko, Zdravko Petkovšek (1965) "Vremenski vodnik za turiste", *Mladinska knjiga*

[2] Leo Breiman (2001) "Random Forests", *Machine Learning Volume 45*, str 5-32

[3] Auroop R. Ganguly, Karsten Steinhaeuser(2008) "Data Mining for Climate Change and Impacts", International Conference on Data Mining Workshops, str. 385-394

[4] Jiawei Han, Micheline Kamber (2006) "Data Mining: Concepts and Techniques", second edition*, Morgan Kaufman Publishers*

[5] Sushilkumar Kalmegh (2015) "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and Random Tree for Classification of Indian News",*International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2,* str. 438-446

[6] Barbara Kastelic (2014) "Analiza neposrednega trženja življenjskih zavarovanj: študija primera", Magistrska naloga, *Fakulteta za management, Univerza na Primorskem*

[7] Igor Kononenko, Marko Robnik Šikonja (2010) "Inteligentni sistemi", *Založba FE in FRI*

[8] Romana Koprivec (2012) "Napovedovanje bolezni ledvic z metodami strojnega učenja", Diplomska naloga, *Fakulteta za računalništvo in informatiko, Univerza v Ljubljani*

[9] Adrian R. Levy, Dov R. Bensimon, Nancy E. Mayo, Henry G. Leighton (1998)"Inclement Weather and the Risk of Hip Fracture", *Epidemiology 9*, str. 172-177

[10] H. Lookman Sithic, T. Balasubramanian (2013) "Survey of Insurance Fraud Detection Using Data Mining Techniques", *International Journal of Innovative Technology and Exploring Engineering* (IJITEE) Volume-2 Issue-3, str. 62-65

[11] John W. Orchard, John W. Powell (2003) "Risk of Knee and Ankle Sprains under Various Weather Conditions in American Football", *Medicine and Science in Sports and Exercise*, str 1118-1123

[12] Andrej Panjan (2009) "Napovedovanje uspešnosti teniških igralcev z metodami strojnega učenja", Diplomska naloga, *Fakulteta za računalništvo in informatiko, Univerza v Ljubljani*

[13] Boris Petelin (2014) "Večnivojski usmerjeni grafi za analizo prostorskih podatkov", Doktorska disertacija, *Fakulteta za računalništvo in informatiko, Univerza v Ljubljani*

[14] Thair Nu Phyu (2009) "Survey of Classification Techniques in Data Mining", *Proceedings of the International Multi Conference of Engineers and Computer Scientists Vol I*

[15] S. D. Prestwich, R. Rossi, S. A. Tarim, and B. Hnich (2014) "Mean-Based Error Measures for Intermittent Demand Forecasting", *International Journal of Production Research 52*, str. 6782-6791

[16] Marko Robnik Šikonja, Igor Kononenko (1997) "An adaptation of relief for attribute estimation in regression", *Proceedings of the Fourteenth International Conference*, str. 296-304

[17] Tadej Seme (2014) "Računalniško krmiljena vremenska postaja", Diplomska naloga, *Fakulteta za naravoslovje in informatiko, Univerza v Mariboru*

[18] Shashi Shekhar, Zhe Jiang, Reem Y. Ali, Emre Eftelioglu, Xun Tang, Venkata M.V. Gunturi, Xun Zhou(2015) "Spatiotemporal Data Mining: A Computational Perspective",*International Journal of Geo-Information 4*, str. 2306-2338

[19] Sebastijan Štraus (2009) "Podatkovno rudarjenje na primeru zavarovalnice Maribor", Diplomska naloga, *Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru*

[20] Špela Valand (2013) "Odkrivanje zavarovalniških goljufij s pomočjo podatkovnega rudarjenja", Magistrska naloga, *Fakulteta za družbene vede, Univerza v Ljubljani*

[21] K. Venkateswara Rao, A. Govardhan, K. V. Chalapati Rao(2012) "Spatiotemporal Data Mining: Issues, Task and Applications", *International Journal of Computer Science & Engineering Survey Vol. 3 No. 1*,str. 39-52

[22] Ian H. Witten, Eibe Frank, Mark A. Hall (2011) "Data Mining: Practical Machine Learning Tools and Techniques", third edition, *Morgan Kaufman Publishers*

[23] Mitja Zupančič, Vera Smole (1997) "Primerjava med kartami fitogeografskih, dialektoloških in etnoloških območij Slovenije", *Traditiones zbornik za slovensko narodopisje in Glasbenonarodopisnega inštituta*, letnik 26, številka 1, str. 49-59

[24] Bernard Ženko (2003)“Izboljšave metode skladanja klasifikatorjev”, Magistrska naloga, *Fakulteta za računalništvo in informatiko, Univerza v Ljubljani*

Viri

[25] The Institution of Engineering and Technology(2015) “Prevention of Slips and Trips”, *Health and Safety Briefing No. 57*

[26] Državna meteorološka služba
Dostopno na: http://meteo.arso.gov.si/

[27] Geodetski inštitut Slovenije
Dostopno na: http://www.gis.si/

[28] Orange - Podatkovno rudarjenje
Dostopno na: http://orange.biolab.si/

[29] Slovenske statistične regije in občine v številkah
Dostopno na: http://www.stat.si/

[30] Slovensko zavarovalno združenje
Dostopno na: http://www.zav-zdruzenje.si/